CHAPTER 19

Bringing Up the System

> ...it is very easy to be blinded to the essential uselessness of them by the sense of achievement you get from getting them to work at all. In other words—and this is the rock-solid foundation on which the [Sirius Cybernetics] Corporation's galaxy-wide success is based—their fundamental design flaws are completely hidden by their superficial design flaws.
>
> —Douglas Adams, *So Long, and Thanks for All the Fish*[†]

19.1 INTRODUCTION

The design work is finished, two-thirds of the electronics seem to run, and though the optical prototype has quite a bit of hot glue on it and lots of scars from engineering changes, the system is over the hump. It's about time to make sure that all the bits connect together properly and actually work in concert. Since the technical risk has been taken care of in the early prototypes, the rest should be fairly routine. The development schedule hasn't slipped much, so from now on it should be 40 hour weeks and lots of time with the kids. System integration and debug is just a matter of turning the crank and fixing small oversights, right?

Usually not. System integration is the time when the really weird bugs show up, the ones that are not reproducible nor susceptible to simulations. It is not that there are evil spirits infesting the subsystems, or that the basic physics is not understood—it's just that it is remarkably difficult to simulate an effect you haven't thought about. Here's a war story.

19.1.1 The Particle Counter That Wouldn't

A group of the author's friends transferred the design for a liquidborne particle counter[‡] to a small vendor. The sensor was to be used to detect submicron foreign particles in fluids (water and aqueous HF primarily). The vendor had previously made dark-field instruments that weren't too complicated. This one was a bright-field Nomarski-type interferometric system with a lot of electronics and software.

[†]Douglas Adams, *So Long, and Thanks for All the Fish*. Harmony Books, New York, 1985.
[‡]J. S. Batchelder and M. A. Taubenblatt, Interferometric detection of forward scattered light from small particles. *Appl. Phys. Lett*. **55**(3), 215–217 (July 1989).

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

The system was complex because it was specially designed to reject spurious counts due to bubbles, the bane of other techniques. Our group had built a few demonstration systems, which successfully operated in the semiconductor manufacturing line, but which were the size of a projection TV, weren't easy to align, and had too many expensive commercial mounts—not a very saleable package. (Still, having a working prototype to start from is pretty comforting.) The vendor had a good optical guy, who undertook to cost-reduce the design for production. Besides designing a package of manageable size and substituting a diode for the HeNe laser, he changed the interferometer by replacing the Nomarski wedges at the pupils of the objectives with Savart plates in front of them.[†] Because of the violently corrosive fluids, the flow cell windows had to be made of sapphire, and to aid servicing, the (very thin) cell had to be removable—a stiff requirement for a moderate-NA system. In addition, they redesigned all the electronics.

Optical particle counters don't really measure the size of particles, but rather their scattering or absorption cross sections. These are converted into fictitious particle diameters by reference to a curve of cross section versus diameter for polystyrene latex spheres, which are a good, commonly available standard. In order to get this cross section from a light pulse, we have to know the intensity of the light incident on the particle. Since optical beams usually have intensity profiles that vary smoothly between the center and the wings, that means we need to know the particle's radial position.

As shown in Figure 19.1, the Nomarski-type interferometer (see Example 10.2) produces two overlapping, orthogonally polarized focused spots, whose phase is adjusted to near 90° in quiescent conditions, so that there is ideally no background signal. Due to the phase shift from the particle, the interferometer generates a positive voltage when the particle is in one of the two spots, and a negative voltage when it is in the other. The particle counter took advantage of this to measure the particle trajectory. By tilting the axis joining the two spots, a particle that crossed closer to the positive beam was made to produce a pulse with a big positive lobe and a weaker negative lobe (and conversely). The signal processing strategy involved measuring the asymmetry of the S-shaped signal pulse, in order to determine where in the beam the particle had passed, so that the extinction and phase shift measurements could be normalized correctly. The details of the pulse shape depended sensitively on the optical system, the laser beam quality, the aberrations in the flow cell window and Savart plates, the state of focus, the impulse response of the amplifiers and filters, and the design of the track/holds in the digitizer, among other things. It was not easy to test these subsystems realistically, especially because the data available from the early prototype was inapplicable due to the large number of engineering changes. In consequence, a great deal was left until the system integration phase.

The first problem was laser mode hopping due to optical feedback. In a dark-field system, we don't usually mind mode hopping much, since the noise is multiplicative—if there's no signal, there's no noise, and a 1% laser fluctuation can only produce a 1% signal level error. In a bright-field system, on the other hand, a 1% fluctuation may completely mask all signals of interest (see Chapter 10 for other ways around this). Although the differential system is nominally zero background due to the subtraction of the two signals, enough of the mode hopping got through to cause serious trouble. This was eventually solved by destroying the temporal coherence of the laser by large

[†]A Savart plate is a symmetrical walkoff plate that shears the two linear polarizations with respect to each other without angular deviation (see Section 6.6.2).

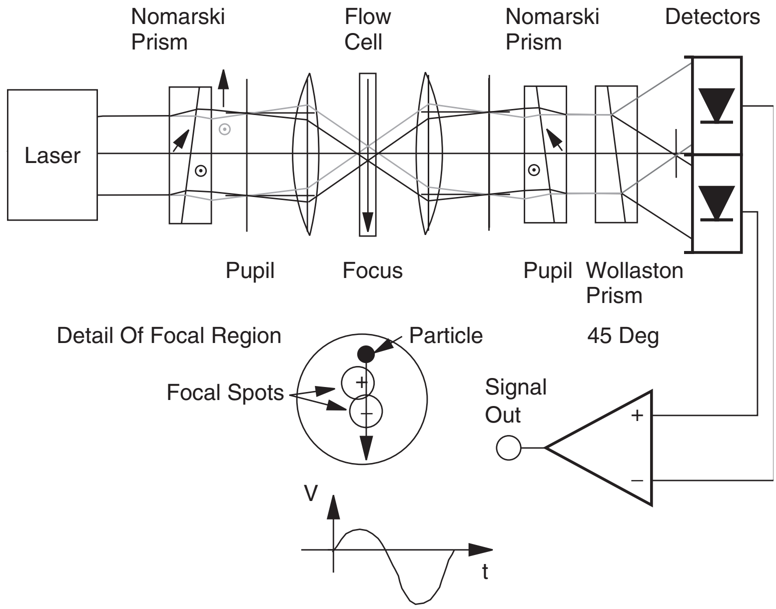

Figure 19.1. Nomarski-type interferometric liquid particle sensor. The Wollaston is oriented at 45° to the Nomarski prisms, so that the two beams are made to interfere losslessly while keeping the background zero.

injection current modulation at UHF (see Section 2.14.3). The coherence length dropped to a few hundred microns, and the mode hopping stopped completely.

If the system had still used Nomarski prisms instead of Savart plates, this would have been a complete solution to the optical part of the problem. Nomarski prisms have very nearly constant phase delay with angle, and at the pupil there is only one plane wave component to worry about. Unfortunately, the delay of a Savart plate is a strong function of field angle and polarization. Besides aberrating the beams, this property interacted with the short coherence length of the UHF modulated diode laser to produce a disaster. We saw in Section 2.5.3 that the relative phase fluctuations of the beams become comparable to a radian at a path difference equal to the coherence length; at 1% of the coherence length, they amount to 0.01 radian. The low frequency fluctuations of the relative phase now became the dominant noise contribution. This was far from simple to solve, eventually requiring a special nonuniform wave plate to correct the relative delay (by this stage, it was no longer trivial to go back to Nomarski prisms).

The flow cell windows were made of *c*-axis normal sapphire (see Section 4.6.1), which looks isotropic near normal incidence. Because the beam was converging, though, this wasn't good enough—window birefringence changed the polarization and state of focus of the marginal rays. This led to a peculiar pulse shape.

The electronics presented other knotty problems, for example, severe thermal drift in a nonlinear amplifier which was implemented with diodes in the feedback loop of an op amp. They were hip-deep in problems. All of them were eventually solved, and the system went on to win a prestigious industry award and get lots of good press, but only after an entire year's schedule slippage and at very serious human cost to the participants. The product line was shortly sold off, and it went out of production soon after.

The force of this tale is that none of their individual design decisions was silly at all—the engineers were a talented bunch of guys, who did most of it right. What killed them was underestimating the difficulty of the system integration and fighting on too many fronts at once.

19.2 AVOIDING CATASTROPHE

19.2.1 Incremental Development

Hindsight is of course an exact science, and in the particle counter case it shows that the best way to proceed would have been to make a much closer copy of the system that was known to work, and to try out each change individually. The system would have incrementally converged on a workable configuration, and the devastating interaction of the optical design changes would have been avoided. Bringing to market a somewhat more expensive but less risky design would have generated revenue, which would have helped to fund the cost reduction steps and so reduce the time pressure.

Horror stories like this can be avoided, once we know what to look for, but most of us have to learn by making the mistake. All practicing electro-optical instrument designers should have the following military maxim taped to their bathroom mirror: "A pint of sweat saves a gallon of blood." Inadequately tested design changes are extremely risky, and it is very easy to underestimate the difficulty when making a change to technologies at the limit of our competence. Very often, such a change seems an attractive way out of our current difficulties; fiber optics and diode lasers especially seem to exert a fatal attraction on designers unfamiliar with their less attractive features (see Section 2.12 and most of Chapter 8). The author once spent a surprisingly long while simply changing from a diode-pumped green Nd:YAG laser to a high power, single longitudinal mode diode laser in an ultrasensitive interferometric system. It took almost a year, off and on, during which time he learned a great deal more about diode lasers than he cared to. It was fortunate that the rest of the system was already working; if the diode had been in the system from the beginning, the design would in all probability have failed altogether. It was only the incremental character of the change that saved it.

19.2.2 Greedy Optimization

> Complicated systems that work are invariably found to have developed from simpler systems that worked.
>
> —Attributed to John von Neumann

Concurrent engineering was a 1990s management fad, but one with some basis: it would be the perfect development strategy if systems engineers were clairvoyant. It extols the virtues of getting a whole lot of people working on a project right from the beginning, with everything being developed at the same time, in order to get a rapid time to market. Since this may involve documenting nonexistent features or designing signal processing for a front end that is misconceived, it is somewhat fraught and requires a durable sense of humor.

This apparently chaotic process is actually very well suited to electro-optical instruments, because it tends to prevent the sorts of disasters recounted earlier. Properly done, engineering of electro-optical systems starts with a conceptual design (see Section 1.1), whose technical risks are carefully assessed and then reduced by breadboarding the hard parts. It proceeds to a detailed system definition and from there, immediately to a brassboard prototype.† This prototype is not pretty to look at, but it does allow every major design decision to be sanity-checked at a very early stage. As subsequent versions of the major subsystems become available, they are immediately tested in concert with the rest of the system, so that integration proceeds in easier steps along the way, and the ability of unconsidered effects to cause catastrophes is greatly reduced. Building prototype hardware also gets everybody asking the right questions—practical ones—and highlights the hard problems. The crucial discipline is, *don't break the prototype*. Ever. Each version change must leave the system working.

Numerical analysts call this strategy *greedy optimization*. We start with a guess and continue to accept each step that makes things better. Numerical greedy algorithms work well with a good enough initial guess, but with a poor one, they can get caught in local minima, or get stuck on a slope that never ends. The incremental engineering model is similar, in that if the initial brassboard is too far from the final system configuration, convergence on an acceptable system may not occur in the time available. Thus it is very important to build a prototype that includes all the fundamental features of the final system. If the final system is to be miniaturized, for example, a breadboard covering a 4 by 10 foot optical table is too far off to be safe.

19.2.3 Specifying the Interfaces

In order to improve our chances of convergence, the confusion must be minimized. Specify all the interfaces in the system in as much detail as possible, in advance, and keep these specifications up to date as the project progresses. As far as possible, subsystems should be independent of each other, so that as long as the interfaces are the same, internal changes to one subsystem do not affect others. Modularity is easiest to achieve in software, followed by electronic hardware, and is most difficult in the optics. This difficulty makes it critically important that a realistic brassboard of the optics be used, starting as early as possible.

19.2.4 Talking to Each Other

Even with incremental development and the use of a brassboard, there is always a very lively possibility that some unnoticed flaw will render the entire system a failure. Reducing that risk is one reason for making sure that the designers and systems engineers talk together a lot, individually and in groups. Get everybody thinking about it by making sure each designer always has an up-to-date copy of the detailed draft system specification, so that oversights can be corrected as early as possible, and that each revision highlights any changes so they're hard to miss.

†Prototypes are named differently, depending on how closely they approximate the final system. In order of increasing sophistication, it goes: mockup/demo, breadboard, brassboard, preproduction.

19.2.5 Rigorous Subsystem Tests

Localize each problem as closely as possible, as early as possible. It is easy to design a system that is the electro-optical equivalent of spaghetti code, or (even worse) of series-string Christmas lights: when one goes out, they all go out. The way to avoid this is by hierarchical design, dividing the function into independently testable subsystems, each of which has its own independently testable blocks. Really work hard at devising good subsystem tests.

19.2.6 Plan the Integration Phase Early

These interface specifications should lead naturally to a plan for system integration; group tightly coupled subsystems together, and test these combinations by themselves, then assemble those few groups into the complete system. This is the same strategy of divide and conquer that is used in troubleshooting: localize the problem as quickly as possible.

A planned system integration is not a substitute for incremental development. If it is used that way, it is nothing but a way of heading for trouble neatly. However, eventually the last prototype version of a system must be left behind, and the jump made to an early production version. Problems must be expected to surface here as well, and this is where the integration plan comes in. Bring up the lowest level of function first—power supplies, computers, light sources. Pay particular attention to the ways in which your subsystem test conditions differ from the actual system environment (e.g., using separate lab power supplies for each electronic board and running the optical system with the covers off). The key to (relatively) painless system integration is to allow lots of time for it, and really watch any places where this system differs significantly from others you've built in the past—removing on-card voltage regulators, switching to diode lasers, changing from dark to bright field (or vice versa), or just building a more complicated system than you're comfortable with can lead to integration problems.

19.2.7 Don't Ship It Till It's Ready

With all we hear nowadays about how time to market is the key to profitability, it is easy to get buffaloed into shipping a system that isn't ready. Missing ship dates is embarrassing; customers and managers get upset; we feel we've let them down by being too slow. Why not ship a couple just to reduce the heat, so we can finish our jobs? They can always be brought up to the final revision level later, after all.

Don't fall for this one. The first customer shipment is a Rubicon that cannot be uncrossed. If you ship an inadequately debugged system, your previously irate customer will now be apoplectic, and so your management will be scared. Scared management does stupid things, such as pouring all available engineering time into customer support, starving the debugging effort, which drags out. Where's your rapid time to market now? Meanwhile, your system is rapidly acquiring an evil reputation among its prospective customers. It is possible to recover from such a disaster, but it is most uncommon (see Figure 19.2).

The one exception to this rule is omitting features. If the basic unit is really solid, but there are one or two features that do not work properly just yet, it is quite permissible to leave them out for the first production run. This applies to the particle counter example, where the omitted features would have been some of the cost-reduction steps. Provided

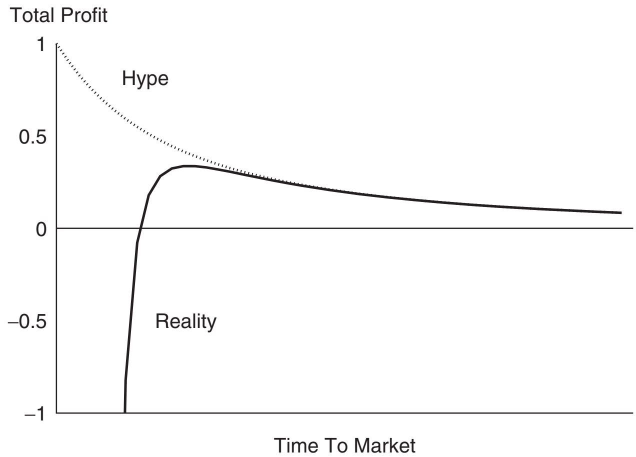

Figure 19.2. Profit versus time to market: hype and reality.

the customer knows just what he will be getting, this can be a valuable source of flexibility in development.

19.3 DEBUGGING AND TROUBLESHOOTING

Troubleshooting and debugging are related but not identical. Troubleshooting is when we're finding out what's wrong with a single device that once worked, or which is supposedly identical to another 23 devices that worked (for arbitrary values of 23). Debugging is getting it to work the first time, or chasing a problem that shows up intermittently in production units.

They can be a blessing or a curse. The blessing comes when the chasing makes us think about the design in a new way, leading to improvements or even to a really new idea. It is also a wonderful intellectual discipline, which designers often need. The curse is when we're unprepared or haven't budgeted time enough.

It's really worthwhile getting good at troubleshooting. Jim Williams, a famous analog circuit designer at Linear Technology, has written an important chapter on fixing things, which says in part:

> The inside of a broken, but well-designed piece of test equipment is an extraordinarily effective classroom. The age or purpose of the instrument is a minor concern. . . . Good design is independent of technology and basically timeless. The clever, elegant, and often interdisciplinary approaches found in many instruments are eye-opening, and frequently directly applicable to your own design work. . . . The specific circuit tricks you see are certainly adaptable and useful, but not nearly as valuable as studying the thought process that produced them.†

Design work has no fixed cutoff point; it's done when the designer says it's done, which is an opportunity for leaving loose ends and unexamined assumptions. There is no

†Jim Williams, ed., *The Art And Science of Analog Circuit Design*. Butterworth-Heinemann, Woburn, MA, 1995, p. 5.

such sloppiness in troubleshooting. Troubleshooting is sometimes thought to be an art, but really it's an applied science. The trouble exists out there somewhere to be found, rather than being created for the occasion; once the bug is successfully found and fixed, everyone can agree that it's gone. (By this criterion design is an art, and debugging has some qualities of both.) All sciences have their own characteristic mindset, and the mindset of a good troubleshooter is a suspicious and very observant one. If you notice anything funny, never, never ignore it just because it doesn't fit with anything else you're working at. Write it down prominently in your notebook (you do keep a notebook, right?) and think about it later.

Debugging is like troubleshooting, only working without a net. This thing has never yet worked right; it's just your confidence that it ought to, and will eventually, that you have to go on. This means that there's no single right answer, and that you can't assume that the design is basically correct. Debugging often uncovers design errors and usually results in design or process changes.

In fact, debugging uncovers many more design mistakes than it does broken pieces, so your initial guess should be that you goofed: not that the seemingly misbehaving Wollaston prism is defective, but that you've got a stray reflection you didn't notice. Most of the construction flaws we find are stupid; for example, the circuit may not be wired the way the schematic says, or a lens may be in backwards. Purely analog electronic debugging is complicated enough that books have been written on it,[†] and when optics, mechanics, digital electronics, and software get mixed in, the result is a real witch's brew.

Frequently, there may be many possible causes for a given problem; for example, say the bias level at the output of the photodiode front end amplifier is drifting. Is this because of circuit board leakage, an oscillating transimpedance stage, etalon fringes, laser instability, or temperature drift? Who knows? You'll just have to make a list of possibilities, then test for them in some intelligently chosen sequence, and with good test gear. If all you have to work with is your own buggy pre-alpha-test software, you'll get turned into a newt for sure.

19.4 GETTING READY

Guiding Principle: Problems *never* "just go away by themselves."

Insist on rational behavior. Circuits are highly interdependent, especially DC-coupled circuits; when one part of the circuit misbehaves, *all* the signal voltages go nuts. In this situation, we have to use what we know about circuits—the component that isn't behaving rationally is the problem. For example, if an op amp's noninverting input is 50 mV above its inverting one, the output had better be railed high. That transistor with the 150 mV V_{BE} had better not be pulling any collector current. Signal voltages shouldn't depend on where along the trace you measure them. Being suspicious and logically scrupulous really pays off.

[†]Robert A. Pease, *Troubleshooting Analog Circuits*. Butterworth-Heinemann, Woburn, MA, 1991.

Divide and Rule. Okay, you have a gizmo that isn't producing the right output. Somewhere inside, something is wrong, but where? The quickest way to find out is to figure out what the major functional blocks in the instrument are, and test somewhere in the middle—misapplying a political analogy, we can call this the *divide-and-rule principle*. If the middle looks right, the problem is in the back half, and if it doesn't, it's in the front half. If the system has a single, well-defined signal path with no ancillary loops such as AGC or APC (automatic phase control, i.e., a PLL), you can do a nice binary search and be sure of finding it in $\log_2 N$ iterations, rather than N if you start from one end and work toward the other. Most systems aren't like that, but generally they're close enough that divide and rule is the right strategy. It's certainly the right approach to traversing our lists of possible problems.

It is frequently useful to cut in at the points where the major transformations occur: optical to electrical, analog to digital, and hardware to software, that is, the detectors, the digitizers, and the bytes read by the computer from the hardware interface. It is easy to look at the signals there, and to inject synthetic "good data."

Examine Your Assumptions. As the saw has it, "logic is an organized way of going wrong with confidence." It's not logic that's at fault, of course, it's that most of us aren't sufficiently rigorous about it. In debugging, we often find that one of our basic assumptions is wrong. This is surprisingly difficult to spot, because of the conceptual blinkers we're wearing, the ones that led us to make the mistake in the first place. The author once spent a long time contemplating an apparently perfectly aligned interferometer that wouldn't work, watching the fringes blink on and off uniformly across the field, before realizing that the beam wasn't well enough collimated, and that since the blinking fringes indicated only the *relative* phase of the two wavefronts, they were insensitive to the collimation error.

If you've eliminated the first thing you thought of, but after a day of head-scratching you can't find what else it could be, go back to the first thing and examine the evidence *very carefully*, preferably explaining it to someone else. Often the problem is hiding there, and your reasoning is what is at fault.

Learn What's Normal. During system integration, a lot of time is spent staring at perfectly functional equipment, trying to figure out what it is doing. Just getting a feel for the normal operating behavior of the system is time consuming and should not be skimped on, so budget time for finding out.

All the design and simulation in the world won't prepare you for what you'll see on the scope. For example, CMOS analog multiplexers exhibit large switching transients when run into high resistance loads, such as a series RC filter. This looks really bad, but in fact the charge injected is always a lot lower than the spec, because the manufacturer's test circuit has a shunt capacitor connected to the switch (as in a T/H). An optical example is learning what defocus or misalignment does to your detected signal, and how to tell them apart.

Believe Your Noise Budget. It is an elementary scientific principle that when theory and experiment disagree irreconcilably, experiment wins. On the other hand, the strength of the evidence needed to overthrow a theory depends on how well established the theory is. In building apparatus, we use well-established theory to tell us how to change the system to get the performance we want, that is, when there's disagreement, *the apparatus* *loses and the theory wins*. Don't discard your noise budget when the system isn't as good as you expected—the purpose of that budget is precisely to help you recognize the problem and fix it. Go through the math again, and think about what physics you might have missed, but don't chicken out and toss it. Electrical engineers seem to be temperamentally prone to this error.

Use Signature Analysis. The advantage a troubleshooter has over a debugger is that he knows that the system can work—it isn't misconceived, no design error will prevent its ever functioning properly. He also has knowledge of how it should work. What the beam quality, bias voltages, and signal levels should be throughout the system have been written down on the drawings, and from familiarity, the troubleshooter learns where to poke the patient to provoke a response.

This knowledge can be systematized into *signature analysis*, which checks the response of the system to carefully chosen stimuli. For example, if you're building a system for detecting Cerenkov light pulses from cosmic ray showers, you don't want to have to wait for a 10^{20} eV proton to arrive in order to check the response of every last PMT amplifier. What you want is a reproducible dim optical stimulus with good timing characteristics, available on demand. This need not be elaborate; in this case, you could use an auto spark plug with a strategically placed series resistor to provide a 5 V pulse for synchronization, plus a few neutral density filters to attenuate the light.

Back ends can be tested by using a digital scope or A/D card to capture a few representative waveforms from the front end, and a D/A card or arbitrary waveform generator (AWG) to replay them into the back end under test, to check for appropriate responses. Make sure that you play them back faithfully enough, or you're just fooling yourself: that means using enough bits, filtering out the clock ripple without group delay distortion of the signal, and making sure the records are long enough for the system to settle down before the event takes place. Test the tests by playing them back into the system they were measured on, and check that the response is identical to what the real signal provoked. Watch out for DC offsets, and remember that fast scopes are good to about 6 bits' resolution.

For each test, a few oscilloscope screen shots pasted into a notebook will enable a good technician to zero in on the location of the problem very rapidly. This is a huge time saver, which is not a negligible consideration when you're building highly complex systems that only the designer may understand completely.

Testing the response to signals in noise may require a white noise generator. Good calibrated ones are available from NoiseCom, and these are especially good for testing thresholding and statistical signal processing systems, for which prerecorded noise is not adequate—you can't measure the detection statistics when the test signals and noise are held in a fixed relationship, as in a recording.

On the other hand, testing statistics-based systems for adequate performance, for example, the very low false alarm rate in the ISICL sensor of Example 1.12, may be very time consuming. For those cases, it's worth having a library of near misses and squeakers-under-the-wire in your recordings, because that way you can verify that the device under test can distinguish them as well as it ought to.

Keep a Good One on Hand. Breadboards of each critical subsystem should be on hand so that suspected bad components can be tested. This works better for electronics than optics. Similarly, a few known good components should be kept around, so that if a bad batch arrives and causes problems, you can pin them down.

19.5 INDISPENSABLE EQUIPMENT

Since debugging and troubleshooting require close reasoning from ambiguous clues, they require first-class data; the reasoning and data have to work together like muscles and bones.

19.5.1 Oscilloscopes

The first thing you absolutely must have if you're going to work on instrument back ends is a good, fast analog scope. Its bandwidth should be at least 350 MHz, and 1 GHz is not too much. Remember that in order to gauge the shape of a waveform at all, you need to be able to see at least the first three harmonics, and to see it accurately, you need the first 10. Even if you're not working at VHF intentionally, many of the components you're using can oscillate up there, and not infrequently they will. SPICE will often miss this, because it depends a lot on layout strays.† If your scope misses it too, finding it will be that much harder.

A sufficiently good digital scope (1 Gs/s) is an adequate substitute, *provided* that it has some sort of envelope or spatial-histogram mode. In those modes, the digitizer works at full speed all the time, and each time bucket on the scope has its upper and lower limits plotted, so your signal is pretty well guaranteed to be somewhere inside that stripe. Live in that mode, and especially avoid "high resolution" modes, which use fast sampling speeds and some simple digital filter to get the illusion of more bits. This is useful occasionally but is seriously misleading in general. Some scopes make you use high res mode for measuring rise times and so on, unfortunately.

Get a scope from a manufacturer that makes really good vertical amplifiers. It is surprising how many scope makers will tout a 1 Gs/s scope that has a 300 MHz vertical bandwidth (for a while, one maker was selling a 2 Gs/s unit with a 350 MHz bandwidth). When called on this one, they will often claim that it is necessary for anti-aliasing protection, but that's just smoke and mirrors—real anti-aliasing would require that the vertical amplifier roll off by at least 40 dB by the folding frequency, and that they do not. Pay special attention to the scope's user interface; a lot of the low end models of the fancy brands are actually rebadged imports, some of which have user interfaces as bad as a $35 DVD player's.

Regardless of your budget, don't buy a superfast scope (>2 GHz bandwidth) unless you really need it. You lose the option of high input impedance, and you trade off an enormous amount in accuracy and edge fidelity to get the extra speed. If you need the bandwidth, sampling scopes such as the Tektronix 11801C (widely available used) have much cleaner edges and are much faster, if you have repetitive signals. (They're also an order of magnitude cheaper.)

19.5.2 Spectrum Analyzers

For any high frequency work, a spectrum analyzer is also a near-necessity. You don't have to have your own private one, but for any sort of serious analog development, you'll

†If you're doing time-domain analysis of a circuit with a high frequency instability, the results will depend sensitively on the time step chosen and will be uniformly completely wrong. Gear-type integrators will mask the high frequency instability and are less sensitive to step size, but their results will be equally wrong.

at least need to have one for every three designers, and be ready to rent some more during system debug. Measurements of close-in modulation and noise on sine wave signals is only possible with an analyzer with a frequency-synthesized LO, so look for that. An analyzer with no frequency synthesizer has to phase lock to your signal in order to work at high resolution, and the phase locking masks frequency drift and close-in phase noise in your signal—very puzzling if you haven't tumbled to it.

Remember that weird problems are often traceable to VHF or even gigahertz oscillations, so get an analyzer that works up to at least 2 GHz. Tracking generators are nice for aligning filters and measuring frequency responses, and with an external directional coupler, they let you use your spectrum analyzer as a scalar network analyzer. Avoid the Swiss Army Knife combination spectrum analyzer/network analyzers, though. The best ones work about as well as amphibious cars, and the worst more like a combination shoe horn and pistol.

19.5.3 Probes

An often-overlooked necessity is good probes. Get several ×10 probes (old time circuit design columnist Tom Kneitel used to say that they were marked 10:1 "because they ten to one away when you're not looking"). All those little grounding doohickeys are important, too. Don't throw them away, and make sure they stay together with the probe. A good, short probe ground is one of the keys to good measurements. Whatever you do, resist the temptation to use long floppy grounds. In the context of a 50 MHz measurement, a 1 inch ground lead or coax pigtail is long and floppy (it's about 20 nH, or 6 Ω). If you like using the 3 inch grounding clip leads that come with the scope probe, reduce the loop area—and hence the ground inductance—by twisting the wire back on itself (like a tangled phone cord, *not* coiled like a spring).

A typical ×10 probe has an input impedance of 10 MΩ in parallel with 10 pF. That sounds great, until you figure out that at 20 MHz, 10 pF has an impedance of 800 Ω, and even less at higher frequencies. At VHF (30–300 MHz) and above, a better probe is often a 450 Ω chip resistor soldered in series with the center conductor of a piece of double-shielded 50 Ω coax, right at the node to be tested, together with a 50 Ω scope input. For a reusable test point, put it in series with an SMB jack (with a good ground).

Sometimes you need a fast probe, but 500 Ω isn't enough. For that case, you can get active probes, which have fast FET amplifiers similar to those used in the scope front end, but mounted right down near the probe tip. Tektronix and Agilent (Hewlett-Packard) make nice ones. You can make your own if gain flatness isn't much of a worry, for example, in a narrowband measurement. These probes are somewhat delicate with respect to overvoltage and electrostatic discharge, but they have 50 Ω outputs, which means you can use them with spectrum analyzers as well as oscilloscopes.

19.6 ANALOG ELECTRONIC TROUBLESHOOTING

One of the places that people get hung up fairly often is in debugging the analog signal processing system. Pease has written a good book on the troubleshooting part of this, and we've discussed ground problems in some detail in Section 16.5.2, so we'll touch on some of the other major debugging issues: pickup, supply coupling, oscillations, and assorted head scratchers such as analog multiplexers that won't switch.

Look for Trouble in the Right Domain. Oscilloscopes are famous for being bad at detecting frequency-domain problems, such as spurious signals or mild distortion, which any spectrum analyzer will pick up right away. On the other hand, frequency-domain instruments often miss things like major-carry glitches in DDSs and DACs—they look fine on a spectrum analyzer, because the glitch energy spreads out over a wide band—it isn't always in the same place in the cycle, so it's not just harmonics. Make sure you look in both domains.

Check that your noise floor is where you expect it to be. This is especially important with sampled data, where aperture jitter causes wideband phase noise.

Don't Expect Inputs to Be Just Inputs. Instruments are designed to tell us about our circuit without adding artifacts of their own. Sometimes, though, they aren't quite as innocent as they look. It is not unusual for a DVM, A/D board, or FFT spectrum analyzer to put small amounts of noise out on its inputs. Because this isn't what we expect to find, it's hard to spot the first few times, and will drive you nuts looking for your mistake, when it's really someone else's. Whenever you get a new instrument, connect its input to a scope and then to a spectrum analyzer, just to see what's coming out.

Don't Use Zillions of Cables. Physicists especially are prone to build apparatus with more coax patch cords than have ever been seen in one place before. This is a guaranteed way to make a measurement flaky. Lab patch cords have short and difficult lives—if you have 20 of them in one place, at least one is broken and two out of the 40 connectors are dirty, bent, or have had their center pin driven out of position. Even new cables talk to one another—don't expect perfect isolation from RG-58. Try to build your prototype signal processing in big enough chunks that you don't need so many cables, and use barrel connectors to join connectorized mixers, filters, and so on.

Get Rid of Pickup. Chasing spurious signals due to stray capacitance and inductance and supply coupling is a nearly inescapable debugging task.

Example 19.1: Unusually Severe Pickup Problems. One time, the author was debugging a card for an inexpensive optical sensor, which used three solar cells and a shadow mask to sense the position of an object in three dimensions (see Figure 19.3). Illumination was provided by IR LEDs IRED1–IRED5, which are chopped at about 100 kHz. This one turned out to be a real onion problem. The new surface mount card was exhibiting severe 100 kHz pickup problems, which nothing seemed to improve.

Onion Layer 1. After a bit of puzzling, it emerged that the 100 kHz was entering from two sources, capacitive pickup and power supply fuzz, which were roughly equal in amplitude.

Because of the femtofarad-level coupling capacitance, the pickup had a relative phase of $+90^\circ$, whereas because of the supply bypass capacitors, the ripple had a phase of -90°; the two were thus 180° out of phase with each other, and because they were nearly the same size, they substantially canceled one another. This was a tricky debugging problem, because almost any change (even beneficial ones) disturbed the balance and made the symptoms worse.

Ultimately, the sources were largely separated by grounding the LED string, so that it was on continuously. This of course made the symptoms worse, but because it cleaned

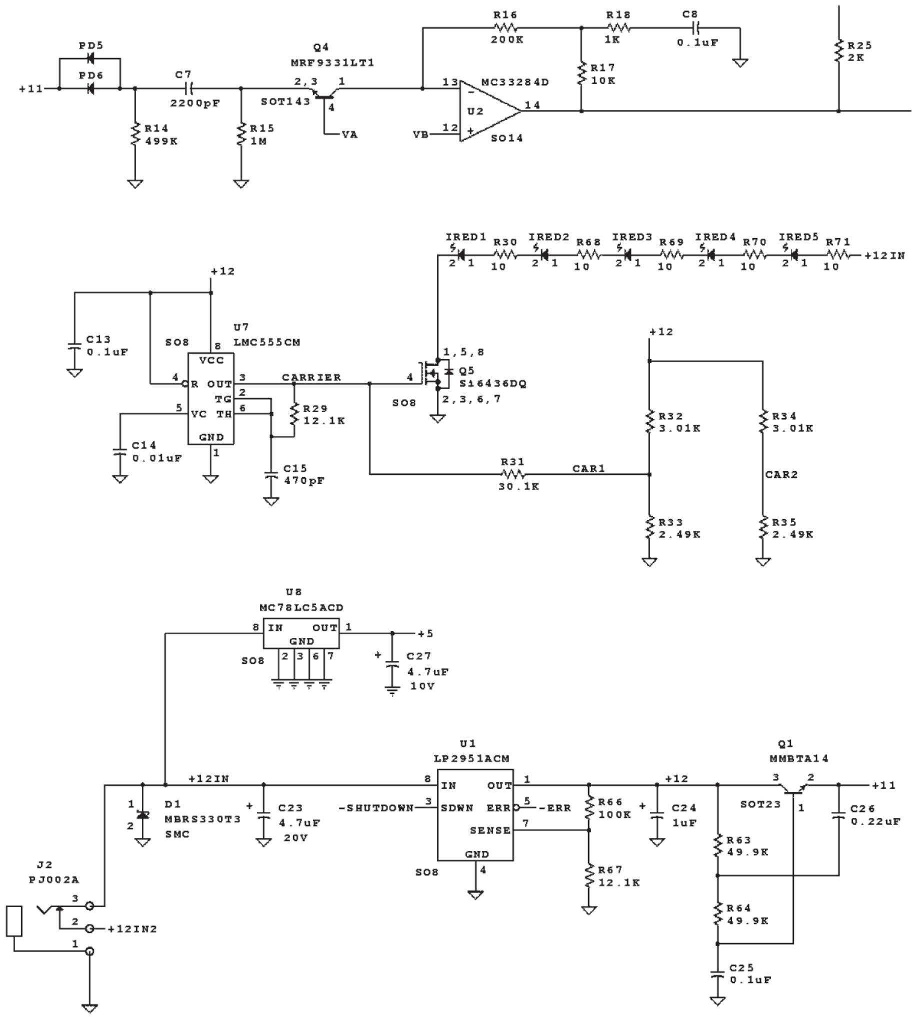

Figure 19.3. Noncontact head tracker schematic fragments: a tiny layout problem caused serious pickup troubles.

up the 250 mV ripple on the 12 V supply somewhat, it seemed like a worthwhile thing to stay with for a while. After cutting traces and hanging wires on two cards to the point where they were too flaky to use, the author applied foil shields to the top and bottom of the cards. This reduced the pickup signal by about 20 dB, which eliminated the cancellation that had been muddying the water. When the LEDs were allowed to switch again, the ripple problem surfaced, but that was cured by putting a 100 μF capacitor from ballast resistor R71 (between the top LED and the supply) to the (grounded) source of the driver FET Q5.

Onion Layer 2. These measures largely cured the problem, except in one channel, where the pickup remained severe enough to rail the digitizer. Since the channels were identical except for layout (all the amplifiers were on the same chip), this looked like a board layout problem. The layout had been done carefully, with ground current loops kept small, analog and digital devices kept separate, and the sensitive analog circuitry laid out one side of the card, over top of the analog ground plane. The power plane was on the far side of the ground plane, split between the ugly +12 V supply under the digital stuff and the +11.2 V analog supply under the analog stuff. Upon closely inspecting the layout once more, the only salient difference was that one single PC pad, 0.8 mm square with a via hole, connected to the sensitive input node of that channel, sat over the singly regulated 11.2 V supply.

This supply was isolated from the ugly unregulated 12 V supply by a low-dropout linear voltage regulator IC (U1), and was used for the less sensitive analog functions. The CMOS 555 timer (U7) that drove the LED driver transistor Q5 was connected to this supply and made it bounce up and down by about 50 mV. This presented no problem for the other two channels, because the front end circuitry was isolated from it by a two-pole Sallen and Key active filter (a fancy capacitance multiplier, see Example 14.1). Since no sensitive traces crossed the 11.2 V plane, it seemed that it should cause no problem to the third channel either. Nonetheless, touching a 1000 μF capacitor from there to ground instantly cured the problem, and permanently moving the V_{cc} pin of the 555 to the 10 Ω–100 μF bypass point completed the cure.

How on earth did this happen? The I-V converter in each channel had a 200 kΩ feedback resistor and had an AC-coupled ×11 voltage divider (see Section 18.8), so that the total first-stage AC transimpedance was 2.2 MΩ. These were followed by phase-sensitive stages (Gilbert cell multipliers), which had a gain of around 400 for in-phase signals (the phase was not particularly carefully tuned—these were primarily frequency converters to allow bandwidth narrowing as in a lock-in amplifier). The size of the detected signal was about 200 mV DC, which at a frequency of 100 kHz means that the required stray capacitance was

$$C_{\text{stray}} = \frac{200\ \text{mV}}{2\pi f V_{\text{pk}} R_f A_V \cos\phi} \approx 16\ \text{fF}, \tag{19.1}$$

assuming that the phase shift ϕ is zero (reasonable, because a 90° lag from the filter capacitor is followed by a 90° lead from the capacitive pickup).

Considering that FR-4 circuit board has a dielectric constant of about 4.5, and that the pad was about 0.020 inch (0.8 mm) from the power plane, the area required to get 16 femtofarads of capacitance is only

$$A_{\text{min}} \approx \frac{4\pi d C_{\text{stray}}(\text{pF})}{1.12\varepsilon_R} \approx 0.003\ \text{cm}^2, \tag{19.2}$$

where fringing has been neglected, all spatial dimensions are in centimeters, and C is in picofarads (the Gaussian unit of capacitance is the centimeter, which is equal to 1.12 pF).

Fringing decreases this number, and the relief of the ground plane under the pad increases it, so this is not an accurate estimate. Nevertheless, it's clear that no very large area is needed—the 0.8 mm pad is about 0.0064 cm^2, so this is an easily believable source of the stray coupling—but only in hindsight. How can this sort of problem be prevented?

In this case, the problem was overconfidence. A dead bug prototype had worked well, and the board layout looked sensible, so the step of estimating the sensitivity numerically was omitted. Once that 16 femtofarad number surfaced, it became obvious that very, very small layout problems could be responsible. In a high gain, high impedance circuit, it is absolutely vital to estimate just how large the stray coupling is allowed to be.

Moral: Greedy optimization isn't always the right approach. For pickup, fix anything you find, regardless of what it appears to do to the symptoms; watch, though, that it is really fixed.

Eliminate Supply Coupling. This example is somewhat unusual in that the coupling was indirect; most of the junk on the supplies had been removed by regulators and active filters. The best way of eliminating supply coupling in general is to make sure the local decoupling of all high level signals and high gain amplifiers is adequate. Note that this does not necessarily mean sprinkling 0.1 μF capacitors everywhere. The key to decoupling is to make the undesired AC path through the supply much higher impedance than the desired one (through the bypass capacitor to the other end of the AC load). This requires locally raising the impedance of the supply with series resistors or low-Q inductors (e.g., ferrite beads), and reducing the impedance of the bypass by choosing the right value and the right capacitor type, or alternatively by using active supply filters, as here.

Sprinkle Active Supply Filters Strategically. In the previous example, after the first onion layer was peeled off (pickup and supply ripple), the 555 was still putting large (50 mV p-p) 100 kHz transients on the 11.2 V supply, which is where the 10 V photodiode bias came from. Even in the face of that 50 mV ripple, and with photodiode capacitances of about 200 pF, it was not necessary to put any additional decoupling on the +10 V supply from the active supply filter. From Eq. (19.1), and noting that the total offset voltage in the well-behaved channels was less than 10 mV, we can estimate that the in-phase spurious signal on the 10 V supply had to be less than 0.2 μV. The calculated rejection of the Sallen and Key Butterworth filter is approximately 140 dB at 100 kHz, so although it is probably operating far less well than this, essentially no ripple at the operating frequency is getting through. This is rather remarkable, because of the apparently direct connection of the quiet 10 V rail to the troublesome +11.2 V; that 5-cent MMBTA14 does a terrific job.

19.7 OSCILLATIONS

Many of the more vexing flaky problems in analog electronics come from high frequency oscillations. Somewhere in our signal processing chain, an emitter follower has a source impedance that is too low, together with a capacitive load, and it starts shrieking at umpty-ump megahertz. Bandwidth limitations in subsequent stages filter out the VHF energy, leaving only the gross bias shifts, intermodulation, phase funnies, and extreme flakiness behind.

Troubleshooting oscillations is easier if you have good test equipment: a 1 GHz oscilloscope, a good spectrum analyzer, and some FET probes. A big, rich lab will have these things, if you can get time on them. In most labs, someone else is usually using them, and you'll have to wait till after hours. Or maybe you don't have access to them at all. In any case, you need some practical skills in troubleshooting them.

19.7.1 My Op Amp Rings at 1 MHz When I Put This Cable on It

We looked at op amp instabilities in Section 15.4.1; they can be caused by too much capacitance on the summing junction or on the output, both of which add excess phase shift to the loop gain, and so reduce the phase margin. Unterminated coaxial cables look like capacitors of 100 pF per meter of length, for frequencies $f \ll 1/(2\pi t_{\text{transit}})$. This capacitance combines with the open-loop output impedance of the amplifier, forming an *RC* lowpass network in series with the feedback path. The phase shift associated with this network reduces the phase margin of the amplifier and produces ringing or even oscillation. It can be corrected by putting a resistor of about the same size or slightly larger than the op amp's open-loop output resistance (look for it in the data sheet or device schematic, though you may have to dig a bit) in series with the output. You can close the DC feedback to the cable's input, as long as you put a small capacitor from the op amp's output to the summing junction. (Note: amps with rail-to-rail outputs are the worst for this.)

19.7.2 When I Wave at It, It Waves Back

If the circuit's behavior depends on where you wave your hands, something is probably oscillating, or your bypassing and shielding is inadequate. It's pretty understandable: we're so often putting a wideband front end amplifier on some device (e.g., a PMT or packaged photodiode) that looks as though its maker intended it only for electric-eye doorbells. We wind up having to ground the cases of photodiodes with copper tape and solder in order to prevent them from picking up FM stations, or oscillating due to input–output coupling. If sensitive high frequency amplifiers are involved, give lots of thought on how to stitch all the grounds together adequately—usually two ground planes with lots of plated-through holes, and perhaps a metal clamp for the diode package.

19.7.3 My Circuit Works Until I Let Go of It

If the circuit is poorly isolated from outside influences, those influences behave like circuit elements. Sometimes these show up as performance variations that depend on where you hold the box, or whether there's a metal surface nearby. These kinds of problems are generically called *hand capacity*. There is a very old and probably apocryphal IBM story of a then-new mainframe that wouldn't work unless there was a scope probe hung from one particular circuit node—the first unit was allegedly shipped like that, and quietly fixed during a preventive maintenance call.

19.7.4 My Transistor Amplifier Oscillates at 100 MHz

Or 1.5 GHz, or someplace that's invisible on your scope. Transistors can usually find something around to use for a resonator, so you'll have to de-Q the oscillation by putting a damping resistor in somewhere. A 10 Ω resistor in series with the base will usually do it. In a differential amp, check if it's a differential or common-mode oscillation, because the medicine is different; a differential oscillation has to be fixed with emitter resistors, whereas a common-mode one usually goes away with a series *RC* (small R, really small C) from the emitters to ground. To check which it is, you can use two probes, or put a 1:1 resistive divider between the collectors and look at the midpoint to see if it moves.

19.7.5 Another Kind of Digital Troubleshooting

The first thing to do with a circuit that is oscillating or has hand capacity problems is to start poking your fingers into it. Not figuratively, either. Touch the input and output of each stage with your finger, while looking to see if the strange behavior changes. (As long as there are no hazardous voltages present—don't try this on an APD preamp!)

Putting your finger on the summing junction (− input) of a fast op amp will cause oscillations rather than cure them, but for discrete stages or noninverting inputs, this test is both easy and sensitive. At RF, human fingers look like a lossy capacitor, which is an excellent temporary fix for most oscillations.

If the finger trick doesn't locate the problem, start looking for impossible bias conditions: things like an emitter follower that is obviously conducting but whose DC bias is impossible (e.g., $V_{BE} = 0.11$ V). That sort of thing can happen during large-signal (Class C) oscillations, in which the transistor is biased off for most of the cycle. Gigahertz oscillations can be really funny that way, for example, an oscillating common-base stage whose apparent V_{BE} changed from 0.15 to 0.4 V when the scope probe was moved a few millimeters along the lead—it looked like a nice DC level on a 500 MHz scope. (It was a biased cascode for a photodiode—see Section 18.4.6.) Remember that (as Pease quotes), "if you notice something funny, Write Down Amount of Funny."

Warning: If you're using a DVM for this, put a balanced decoupling network consisting of a 10 kΩ resistor in series with each lead, and a 10 nF capacitor connecting them on the meter side. Large amounts of RF will drive most DVMs *crazy*, not to mention what those nice antenna-like leads will do to the circuit—if it wasn't oscillating before, it probably will when you hang a meter of probe wire directly on it. Once you've found the stage that is most sensitive, you can try the following tricks.

Add 10–100 Ω in the base lead.

Add extra supply decoupling—1–20Ω in the supply lead.

Put ferrite beads on base and emitter.

Put 10 Ω in series with the load. Normally, putting 10–100 Ω in any two leads of a BJT will stop UHF oscillations.

Put an *RC* snubber on the output.

Use constant-resistance diplexers. This eliminates the reactive load.

Reduce the collector current. This reduces f_T and hence the high frequency gain.

Use a slower device.

19.8 OTHER COMMON PROBLEMS

In debugging and integration, remember that both are onion problems: once the worst problem has been fixed, the next worst will surface, and so on until we get exhausted or the system is perfect. If something we do affects a problem, that usually means that we're getting close to its cause; often, being able to make the problem worse is nearly as valuable as being able to improve it.

Hmm. Must Be a Dead Chip. . . . This is a strong possibility in troubleshooting, but not in debugging, unless it was working before, and you just destroyed it, for example, by putting it in its socket backwards and applying power (if you did that, don't even bother checking—chuck it). It is amazingly rare to get a bad part from a reputable supplier these days. If the circuit hasn't worked yet, it's almost certainly a design or wiring error. Check all the supply voltages on all supply pins of all ICs. Check that you know why any unconnected pin is unconnected—usually because it's an unused output or an offset adjust. Anything else needs to go somewhere, especially resets, chip enables, three-state enables, and so on.

This Precision Op Amp Has an Offset of 100 mV. Assuming the chip is good (which we always do at first, right?) and that you haven't done anything perverse with the offset adjust pins, the main possibilities here are exceeding the input common-mode range, out-of-spec supply voltages, nonlinearity due to slew limiting, saturation, or oscillation, or a DC error due to the input bias current times a huge source resistance. You see this symptom often with high-Q active filters and the early stages of a signal processing system—subsequent filters mask the clipping, leaving only the low frequency nonlinearity, offsets, and splatter—signals showing up frequencies where they have no business being.

Sometimes I Have to Cycle Power to Get It to Work. Some circuits work fine once they're going, but have trouble starting up properly when you power them on. Oscillators are sometimes like this, when the large-signal and small-signal behavior of the transistors are very different. Startup problems are especially likely when you have cascaded NPN and PNP differential pairs, or op amps whose output swing is bigger than their input common-mode voltage range, because in those situations there are two stable operating points: besides normal bias, you might have the output wake up at one rail, which will completely turn off the input stage of the following amplifier. Depending on the design, this may be harmless and self-correcting, or it might destroy the whole system (as in a motor control loop). Make sure you consider what happens under all imaginable startup conditions, with one supply coming up before the other or both coming up slowly, or one pausing and going down again for a short time, and so on. Startup behavior is usually different at different temperatures and supply voltages, so make sure you cover the whole range. Startup circuits and power on resets really are surprisingly subtle.

It Can't Handle the Input. Overdriving IC inputs can lead to very strange behavior, for example, in CMOS, you get charge migrating all around the isolation tubs and changing all the threshold voltages, and some op amps reverse their output polarity when their inputs are driven too close to the positive supply, which can lead to the destruction of your motorized or temperature-controlled device. Watch out for lockup conditions, and always test short circuit all inputs and outputs to their supplies and grounds (make sure you design things to survive this). Some op amps can drive their own inputs nuts—especially watch CMOS ones with rail-to-rail outputs but not rail-to-rail inputs. Instruments can be saturated by strong out-of-band signals—watch out for oscilloscopes, lock-in amplifiers, and FFT analyzers here.

It Can't Handle the Load. Overloading the output of an amplifier will usually cause worse problems than simply reduced output swing. Symptoms include sudden death (e.g., the function generator with a "50 Ω" output that blew up when 50 Ω was attached, because its output transistors wouldn't handle the current); clipping due to current limiting; frequency response degradation; distortion; or oscillation due to protection circuitry or to capacitive loads.

My Low Noise Amplifier Has a Weird Gain Peak. This is a common problem with low voltage noise op amps when used with high source and feedback impedances ("high" can be 1 kΩ sometimes). There is a big peak in the hundreds of kilohertz to low megahertz, which is hard to explain; it's actually a mild case of the same disease that causes loop oscillations—loading due to input or load capacitance. There's more in Section 18.4.2.

My Simulation Is Perfect, Why Doesn't It Work? Usually because of simulation inaccuracies, circuit strays, and supply coupling. Try putting your finger on the supply trace and see what happens. Improve the grounding and decoupling by massive overkill. Then look again—it'll probably be a lot closer, provided you did a really realistic simulation (i.e., included the actual package and PC board strays, parametric variations of components, transistor-level models of your ICs rather than macromodels, and so forth). Theory works, but circuits are hard to model exactly, especially in the time domain.

I Don't Understand My Oscilloscope Trace; It Doesn't Make Any Sense. . . . If you're using a digital scope, run it up to its fastest horizontal sweep rate and work down. You're probably seeing aliasing. The 20 MHz and 100 MHz input filters (usually on the vertical menu) are often helpful in sorting this out. You ought to be using the HISTOGRAM or ENVELOPE function, which runs the digitizer at full speed and displays a color histogram (or at least the high and low values) in every pixel width. Use the "Sample" mode only for automated measurements (e.g., rise time) and absolutely avoid the "High Res" mode.

My Entire Analog Multiplexer Circuit Wigged Out. A seldom-discussed pathology of CMOS circuits is their tendency to connect all their pins together when even one is overdriven. This weird behavior is caused by the input protection circuitry, among other things, and happens when any pin is driven beyond the supply or ground rail. If a circuit containing a multiplexer or analog switch IC is acting strangely, failing to switch properly, and all the inputs seem to be loaded down in unaccountable fashion, look carefully for somebody being driven beyond the supply.

I'm Dying of Capacitance. A common complaint, unfortunately. Low level, high impedance signals have to be shielded to protect them from picking up TV signals and so on, but coaxial cable has a capacitance of 100 pF per meter, which rolls the bandwidth off really badly. There are three basic techniques for this. The best is to put the first-stage amplifier right at the transducer (photodiode, microphone, what have you). This is usually possible with a bit of creativity; you can put a hermetically sealed (ceramic or metal can) op amp inside a vacuum chamber, if you heat-sink it to the chamber walls (you'll probably have to drill a hole in it for UHV use, and watch out for soft solder). Next best is reducing the input impedance with a really quiet common-base input stage or transimpedance amp, as we did in Chapter 18. (Remember what that did to the input voltage noise of the amplifier, though.) The neatest-looking but least satisfactory method is guarding, that is, driving the cable shield from the output of a voltage follower, to bootstrap the cable capacitance. In Section 18.4.8 we saw that bootstrapping can work well but has the same noise-multiplying tendency as transimpedance amps; your frequency response will improve, but your SNR won't.

My Network Won't Tune. In Section 16.11, we did our tuning by a greedy algorithm—tune knob #1 to a return loss maximum, switch to knob #2, and continue round robin fashion—and discussed what to do in noisy situations where the indicators are unreliable: tune in big enough steps that you can clearly see where things are getting worse, and head for the middle.

My Network Won't Tune, Part 2. There are other reasons for networks failing to tune. In a matching network, the maximum impedance transformation ratio is Q^2, so if your circuit Q is too low, the matching condition will be inaccessible. In that case, you can increase the Q by making your series inductors and shunt capacitors bigger, and your shunt inductors and series capacitors smaller, then try again. Calculate first, though, because that's a fair amount of work; keeping the resonant frequency of each section constant in the process is usually enough. Maybe the most common problem of all is running out of range. There should be two equivalent positions of each trimmer capacitor—if not, you've run out of range.

My DC Level Is Drifting All Over the Place. Watch out for thermoelectric potentials, especially gradients across active devices. Thermoelectric coefficients of wire thermocouples are usually 20–50 μV/°C, but Cu–Si's is 400 μV/°C, and copper/tin oxide is even bigger. Soldered connections are not usually a problem, but Cd–Sn solder has a bit lower thermoelectric coefficient with copper than Pb–Sn has, and Cu–Ag and Cu–Au are very low.

Narrowing My Bandwidth Isn't Helping. In Section 13.6.1, we saw that drift and $1/f$ noise isn't helped by bandwidth narrowing alone. You'll have to get more of the measurement bandwidth further from DC, with AC modulation or fast scanning and signal averaging. You might then find that you can widen your bandwidth instead.

My Noise Looks Like Data. Some measurements, especially slow-scanning ones like magnetic force microscopy, are prone to generate reasonable-looking wrong answers. This is a difficult test of one's scientific integrity sometimes; it helps to remember that other people (customers or readers of your paper) are going to be ruthless about examining it, so you might as well be too. If you wouldn't believe it in somebody else's work, where you weren't too familiar with the apparatus and had no personal stake, others are unlikely to believe yours either. This is a tough one; some people are their own severest critics, so apply this rule with some sensitivity, and show the data to a friendly colleague.

19.9 DEBUGGING AND TROUBLESHOOTING OPTICAL SUBSYSTEMS

There's a considerable amount of overlap between this topic and Chapter 12, because unlike electronics, just assembling an optical system requires a lot of testing and checking as we go along. We'll concentrate here on properly aligned systems that don't work properly.

Optical debugging is simpler than electronic in one way, because there are normally very many fewer optical elements than electronic ones. In other ways it is more complicated, because each element does much more complicated things, and the available test equipment is impoverished compared with oscilloscopes and spectrum analyzers. It isn't too bad in the visible, where you can see where the beam got vignetted (say), but in the IR and UV, it isn't so easy.

This difficulty is aggravated by any unclarity in our intuitive idea of just what the system does, and how. It is crucially important in optical debugging to be able to describe the operation of each part of the system in at least two complementary ways, for example, geometric imaging, wave propagation, and photon following. Different things are obvious in different pictures, and thus a flexible understanding is a vital tool.

The most common optical debugging problems are caused by things you didn't think of, for example, that iris-type shutters cause intensity nonuniformity unless they're at the pupil, or that a thin lens illuminated with a laser exhibits very fine Newton's rings. The second most common type comes from having too simple a mental model of how things behave, so you miss their subtle interactions; diode laser feedback and thermal lensing are examples. Table 19.1 has some possible sources of instability to start from.

Here's some advice.

Each Mode Is Its Own Interferometer. Lasers often oscillate in several modes simultaneously, so that their optical spectra look like fine-toothed combs with 2 to 100 teeth. Because the speed of light is so high, the mode spacing in ordinary sized lasers is 100 MHz (in a big gas laser) to >10 GHz in a Fabry–Perot type diode laser. Detection makes all these modes interfere with themselves and each other, producing a strong baseband signal plus a forest of huge ugly spurs near harmonics of the mode spacing, and

TABLE 19.1. Some Sources of Instability in Bulk Optical Measurements

Optical
Etalon fringes
Polarization nonuniformity (especially in VCSELs)
Aberrated beams
Misalignment
Not enough space for misalignment to show up clearly
Schlieren due to air currents or temperature gradients
Vignetting
Stray light
Glints
Laser
Collimation errors
Interferometer path difference $\neq 2N$ cavity lengths
Wiggle noise due to spatial side mode
Mode partition noise
Mode hopping
Interference with delayed reflections
Coherence fluctuations
Mechanical
Order overlap in gratings
Bending due to thermal gradients
Mechanical creep
Hysteresis
Backlash
Stick–slip energy storage
Lubricant film instability

(at least in gas lasers) unstable baseband spurs caused by the anharmonicity of the comb (see Section 2.13.7). Because of the spurs, we confine ourselves to the neighborhood of DC, and so for our purposes, each mode is its own interferometer.

There are two important consequences of this: first, an interferometer using a multimode laser must have a path difference that is close to $2N$ cavity lengths (taking the refractive index into account of course), or else the sensitivity will be reduced and the noise increased, since the different phases of the different modes cause mode partition noise to be transformed into intensity noise (see Section 2.5.3). Second, any high frequency components of the desired signal will be mixed down into the baseband region and made unstable by the frequency instability of the mode spurs.† A noisy interferometer can often be traced to one or both of these effects, so dig out a ruler and check that $\Delta z/\ell = 0, 2, 4, \ldots.$

Mentally Follow Your Photons. Our familiar descriptions of wave propagation in terms of plane waves, rays, and δ-functions are all singular in some way, and hence unphysical. A plane wave has infinite energy, so (although the math works out fine) it is a bit unintuitive sometimes to think of waves of finite extent and finite power being sums of plane waves. If it's hard to get a crystal-clear view of how some optical subsystem should behave, try changing domains. Go back and forth between frequency, space, and wave packets (pulses or Gaussian beams). Thinking about a general beam as a superposition of Gaussian or tiled, square uniform beams at different positions is often helpful too, for example, thinking of all the tiled components of a high-NA focused beam not adding up as scalar addition, due to their different propagation angles (Section 1.3.1). (See Figure 19.4.)

Use Different Descriptions. There are often a few different equivalent ways to describe the physics of your setup. For example, consider a system for inspecting a smooth but not necessarily flat transparent film for defects, as shown in Figure 19.5. The system uses a two-pass geometry with a small, fairly dim source and a low tech imaging system, plus a big sheet of cheap retroreflector to send the light back to the source.

The physics can be described as two passes through a pellicle, or as four independent beams (top, top), (top, bottom), (bottom, top), (top, top). Phase and polarization issues and other delicate effects are often more easily understood by switching back and forth this way. Another well-known example is the plane mirror Fabry–Perot interferometer, which can be solved by a geometric series approach as in Section 1.6.2 or by patching together the plane waves inside and outside the cavity.

Watch for Polarization. Polarization funnies can drive you nuts. If you're using a polarized light source, it's worth following the polarization around with an analyzer to make sure it does what you expect, and watch out for sign errors (e.g., right- and left-circular), because that's where a lot of mistakes lie. For example, beyond Brewster's angle a right-handed elliptical polarization stays right-handed on reflection.

Get a few polarizers of good optical quality and wide angular aperture (film for the visible, Polarcor for the near-IR, Rochon prism or walkoff plate for the UV), to make sure that the polarization does what you expect. Mark their orientation clearly, and look

†The modes are remarkably narrow as a fraction of the optical carrier frequency (10^6 Hz/10^{15} Hz), but much wider as a fraction of the mode spacing, which is what's relevant in the baseband signal.

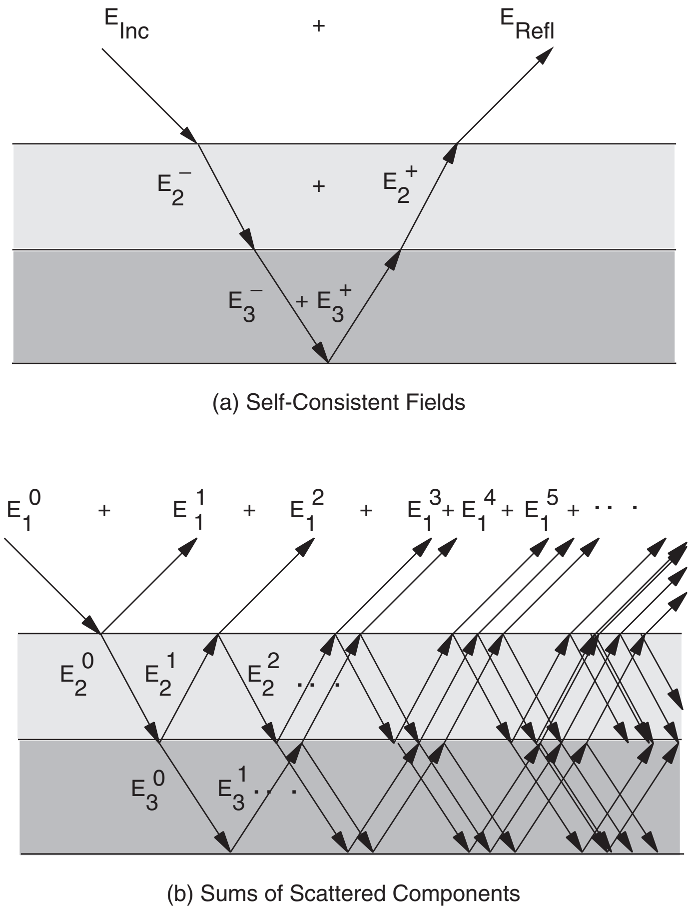

Figure 19.4. Following photons through a thin-film stack: self-consistent fields versus sums of single-path contributions.

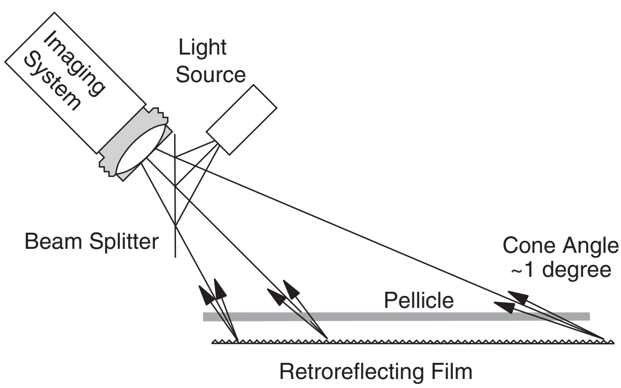

Figure 19.5. Two-pass inspection system for surface defects.

hard; polarization effects are sometimes subtle, and small angular errors can have a big effect on measurements.

Stress birefringence and retardation nonuniformity in wave plates and Pockels cells are frequent villains in systems needing high extinction. A less obvious one is frustrated TIR, for example, from optically contacted surfaces that aren't quite in contact, which can cause really major polarization funnies. These difficulties usually change significantly with temperature, so a little work with a hair dryer and crossed polarizers will often tell a clear story.

Know Where Your Stray Light Goes. Front surface reflections are a bit like land mines: they don't disappear on their own, and the only way to be safe from them is to know exactly where they are. Every dielectric surface produces at least two reflections. Make sure you know exactly where each one goes, and what it will hit. This becomes so complicated after a few bounces that you'll probably fall back on measuring it instead, which is OK as long as it's done in a realistic setting. For example, using a HeNe laser to measure the back-reflection from a laser diode collimator ignores the wildly nonlinear nature of the interaction between the back-reflection and the laser cavity.

First-Order Analysis Isn't the Whole Story. Physicists (e.g., the author) are especially prone to get snookered by first-order analysis, for example, "components above f_c get chopped off by the filter," neglecting the fact that this filter might easily have a shape factor of 30, so that the transition band is very important. Similarly, "polarization insensitive" doesn't mean zero polarization sensitivity, and "shot noise limited" doesn't mean that all your outliers couldn't be coming from an amplifier with popcorn noise, even though its rms contribution is way below the shot noise. The wisdom here is to calculate a hard upper limit to the effect you're neglecting before neglecting it, probably making a few measurements along the way.

Etalon Fringes Are Sensitive to Everything. One of the things that makes optical systems mysterious sometimes is that they're hundreds of thousands of wavelengths across, and often sensitive to effects at the 1 ppm level. Thus parts-per-billion effects can easily become a problem. Gentle air currents, too slow to detect on the skin, can cause big enough index gradients in the air to steer beams around a little: the effect isn't much, but the amplifying effects of etalon fringes can turn it into serious low frequency instability in your measurement. You can often spot this by waving your hand around near the optical path. Small temperature gradients can do the same thing; dn/dT for air is comparable to that of glass. The solution is good packaging and, even better, getting rid of the fringes.

19.10 LOCALIZING THE PROBLEM

If the optical problems are unstable with time, it's an etalon fringe, thermal gradient, or source instability. If your SNR is too low, check for major design blunders, for example, a misplaced factor of 10 or assuming that the focal volume of a Gaussian beam is equal to πw^2 times twice the Rayleigh range (that's a factor of 10 right there[†]), and then look

[†]This one bit the author once, leading eventually to an 8× reduction in the system specification—and that was after having promised the original spec. Very embarrassing.

for places you're losing a decibel at a time. If your beam quality is poor, look for laser feedback. If your CCD pictures are ugly, look for bloom and clock level problems, or too low a temperature in a cooled CCD camera; if that isn't it, put an IR blocking filter in front.

One technique that isn't used enough is intelligent manual manipulation. If your baseline is swimming around all over the place, try pushing with your fingers on the optical mounts, flexing the lens barrels, tapping things with the eraser end of a pencil, breathing on things to warm them up a bit and stir up some index gradients. By finding the place that's most sensitive, you've got much closer to the locus of the problem.

19.10.1 Is It Optical or Electronic?

When chasing some unknown spurious junk in your signal, the first question that has to be answered is whether it's optical or electronic in origin. This is a special case of the divide-and-rule principle but is worth putting a box around because the two cases need such different treatment. Often it's as simple as blocking the light beam and seeing if the junk goes away, but that can be seriously misleading sometimes; for example, consider a front end amplifier that oscillates whenever the photocurrent exceeds 100 μA, as can easily happen with a common-base input stage without a base resistor. That's an entirely electronic problem that goes away when you block the light beam. Other systems (e.g., power-stabilized lasers) won't operate properly without the beam present.

Thus determining which domain the problem lies in can be a subtle business. The first rule is not to actually turn anything off, because doing that changes both the optical and electronic operating conditions, and hence muddies the water. Try attenuating the light gradually, with a polarizer, an ND filter, or slight vignetting, and see if the junk is proportional to the optical power, which suggests an optical problem, or perhaps cuts off at some nonzero beam brightness, which suggests an electrical problem. Note whether it cuts on again at exactly the same place—hysteresis here is a nearly infallible indication of an electronic problem.

Alternatively, replace the light source with a flashlight to maintain the DC bias conditions. Statically balanced systems like dual-beam spectroscopy or DC Nomarski setups can be fooled by replacing one of the beams with a flashlight; autobalanced systems such as laser noise cancelers can be tested by unbalancing the two arms by adding a bit of quiet light to one arm; if the noise gets much worse without changing significantly in character otherwise, the problem is probably optical. If the problem goes away when you turn off the room lights, it's probably optical too.

If none of these tricks is applicable, you can examine the light beam itself to see if you can see the problem there; put in a beam sampler of some sort, and detect the resulting sample, looking for similar spurious signals in the sample output. The main thing to be aware of is that simple-minded tests can be very misleading here—use a few lines of evidence, and keep the possibility that you're wrong in the back of your mind.

19.10.2 Component Tests

Many optical components are made in small batches, with a lot of hand work and without logistics systems of any sophistication. Thus there is some likelihood of a mixup, such as one lot not being tested, or the wrong part shipped. This is sometimes obvious, but often not—and then it will reliably give you fits. One real-life example was half a dozen beam separator cubes[†] in which the quarter-wave plate was tuned to the wrong wavelength, so the transmit/receive selectivity was only 15 dB, rather than >30 dB. This didn't show up until the diode laser started mode hopping madly. It's really worthwhile testing critical optical components before using them. Furthermore, swapping in a new one and seeing if it helps is an unreliable test unless the new one is really a known-good old one, or at least from a different manufacturing batch. In the beam separator case, all six were identical in performance, but identically wrong.

19.10.3 Beam Quality Tests

Intensity ripple and scattered light are common problems, too. A video camera and frame grabber with some simple software to display a false colored image of beam intensity versus position is a big help, but you can also just shine the light on the back wall of the lab; intensity ripple has to show up as far-field artifacts eventually. Knife-edge and chopping tests are a poor match for this, because they integrate over one dimension, which usually washes out the variations we're trying to see.

19.10.4 Collimated Beam Problems

A shear plate[‡] inserted into the beam will show up gross collimation errors or aberrations; it needn't be at 45° to work, so you ought to be able to fit it in somewhere. It isn't particularly sensitive, so for more subtle beam quality problems, use a pickoff mirror and a better-grade collimation tester such as a shear plate dithered back and forth in angle (like the late lamented Collimeter, formerly made by Blue Sky Research—see Section 12.6.4), or a measuring interferometer.

19.10.5 Focused Beam Problems

Test focused beams with a knife edge. A piece of a broken mirror makes a really good knife edge for systems that are already assembled, because it can send the test image off at a convenient angle; it isn't easy to make a good reflecting knife edge any other way. The illuminated area is very small, so straightness of the edge and flatness of the mirror hardly matter—aluminized window glass is fine. You can find the focus in the usual way, by looking at which way the shadow moves when you move the blade, and measure the spot diameter by plotting total reflected light versus blade position. Do make sure the motion is perpendicular to the edge, or you'll be off by a factor of sec θ. You can mount the mirror shard on a piece of bent aluminum sheet hanging from one screw like the blade of a paper cutter, so that it can swing in and out as needed.

If the knife edge test shows a badly aberrated beam, you can often identify the problem by racking the focus in and out and looking at the near-field pattern. Gross astigmatism shows up clearly as well-separated tangential and sagittal foci, for example.

Beam scanning systems can use bar targets, for example, the ubiquitous USAF 1951 three-bar target, which function like a chopper test (the inverse of a knife edge); the

[†]That is, polarizing beamsplitters with $\lambda/4$ plates cemented on one end, for the circular polarized transmit/receive duplexer trick.

[‡]For example, an unwedged optical flat or a specially selected microscope slide that happens to have a sweet spot where the faces are flat and parallel.

beam moves and the blade stands still. The beam usually scans too fast to be examined in detail, but the fact that you're measuring the real signal in a realistic way offsets that disadvantage. That also works much better with invisible light, because it doesn't rely on eyeball estimates. Watch out for asymmetric aberrations here—an astigmatic spot (which is long and thin) shows up beautifully on a properly aligned bar target, but degrades far more rapidly than a symmetrical one as the target is twisted.

19.10.6 Viewing Techniques

In Section 11.8, you were exhorted to put a viewer into the system. That idea can be exploited for debugging and troubleshooting, either by using the built-in viewer and racking focus in and out, or by examining a low-NA aerial focus with a microscope. Astigmatism, coma, and spherical aberration show up pretty clearly this way. Before doing this, calculate the maximum the power that can exit the eyepieces, and after it's installed, measure it to make sure it's safe and comfortable (1–10 μW is a good range, nice and bright without being painful, and far too dim for any danger).

19.10.7 Test Techniques for Imaging Systems

Imaging system performance is best characterized by MTF versus position and wavelength, contrast, uniformity, and efficiency. If you're using a focal plane array detector (e.g., a CCD), some simple testing software and a set of sinusoidal test reticles are a more or less complete solution for detailed testing, but during setup we need something quicker, that we can eyeball. One good method is to use a radial bar pattern, and mark it with a sharp-tipped felt pen and a circle template, to indicate the finest lines your system should be able to reproduce. If you're handy with Postscript, you can generate reticles like this on paper, on demand, which is pretty convenient.

19.10.8 Test Techniques for Light Buckets

Systems with little or no spatial resolution (e.g., photometers and differential detectors) usually rely heavily on stability, linearity, and good baffles. Long-term stability is hard to measure really adequately, especially during troubleshooting, but surrogates exist. Intensity instability is usually caused by thermal drift of filters, etalon fringes, and the leakage current of detectors and front end amplifiers. Warming with a hair dryer is a good test for this, but watch out for Schlieren effects from air turbulence—warm the system up and watch it as it cools down undisturbed. Putting in a chopper is another good test in a DC system—detector and amplifier leakage won't get translated to AC, whereas optical instability will; comparing the AC and DC outputs will identify the source of the trouble. The easiest way to check for linearity in a light bucket system is to use an LED with a small audio-frequency sinusoidal modulation on top of a constant DC drive current. Move the LED around or attenuate it with ND filters, while looking at the ratio of the AC and DC signals on a couple of DVMs—in a really linear system, the two outputs will track exactly. Alternatively, two LEDs in an integrating sphere, with different modulation, will generate mixing products in a nonlinear system. You need the sphere to keep the intensity ratio independent of position.

19.10.9 Invisible Light

All these tests get harder in the UV and IR. Inexpensive cameras work between about 250 nm and 1 μ m, but outside that range things get difficult. Fluor-coated CCDs work well into the UV, and an image tube IR viewer with an S-1 photocathode gets to 1.3 μ m for really bright light. Further into the IR, life is really miserable. Lead-salt vidicon cameras get to 2.2 μm but have such strong memory effects that they are nearly useless for fine work. Mid-IR focal plane array cameras are very expensive, although the new bolometer array units are beginning to change this.

If you do use a camera, get an articulated mount such as a Bogen Magic Arm to hold it. Magic Arms are beautiful—they're about the size of your own arm, and articulate the same way, but can be twisted into any position and orientation and then frozen solid by 1/4 turn of a lever. They have 1/4-20 mounting holes, so they bolt right onto your optical table, and also come with an excellent clamp. Cameras are bad enough without adding pointing problems.

19.10.10 Test Techniques for Fiber Systems

There are fewer things to go wrong with fiber systems, but on the other hand, they are more pervasive and much harder to fix. Trouble from cladding modes can be eliminated by mandrel wrapping or potting the stripped fiber end in wax. Polarization problems you just have to live with, although switching to polarization-maintaining fiber helps a fair amount. Galloping etalon fringes are a sad fact of life, that only Faraday isolators plus angled fiber ends, or very low coherence sources can fix (the latter, at the price of FM–AM conversion). Broken or dirty fibers are pretty obvious, except in bundles; collimation errors you fix as usual. Microbending loss can be due to too-short wavelengths where there is a leaky mode, or to a poor mounting scheme (e.g., metal gooseneck armor).

The major effects of thermal and mechanical instabilities on fiber systems are excess loss in launching light into the fiber, microphonic effects due to wiggling launchers and acoustic microbending and fringes of a few percent P-V going by constantly.

If your launching efficiency is poor, look for defocus, misalignment, dirt, surface damage, or a bad cleave. Cleave problems include hackles (little chunks added or missing where the crack front took a sharp turn) and undesired facet tilts, which misdirect the light. A fiber microscope is an easy way to look for this. In connectorized systems, the fiber might have pulled out or not have been polished well enough; dust on a physical-contact connector can destroy the facet.

19.10.11 Test Techniques for Frequency-Selective Systems

The main things that go wrong with grating systems are misalignment, ghosts, and stray light. Misalignment shows up in broadened or asymmetric lines, low efficiency, and wavelength variations along the exit slit. Use a narrowband source and look for the light intensity peak in the output slit moving around with slight detuning. To look for stray light, use a filtered narrowband source, detune the instrument, and detect with a photomultiplier. The amount of stray light usually depends strongly on the illumination angle; it gets worse when you fall off the grating.

Colored glass filters tend to be fluorescent, and some will drift with time, temperature, and abuse, so try changing their order or rotating them to see if anything changes; the desired signal should not change at all, because the transmittance of a string of filters should be the product of their individual transmittances.

19.10.12 Source Noise Problems

Laser noise always varies with position and angle. Try putting a black-painted razor blade, a bit of cleaved silicon, or a shard of broken mirror on a translation stage, and gradually vignette the beam while looking at the detected noise level—often it'll be nearly a minimum with an unvignetted beam and no sample, but get big when you cut off half the beam. Try it from a couple of different directions, too. Gas lasers with relatively poor mode selectivity (e.g., big argons) are especially prone to this; a spatial side mode that is close to oscillating will make the beam wobble around very badly. (Pick a knife edge that minimizes the scattered light, to avoid confusion from speckle or mode hopping.)

Side-mode problems in laser diodes are easier to spot, because the mode separation is big enough that you can use a grating. Try coming out of the grating near grazing incidence; the anamorphic reduction in beam width (and attendant angular spreading) makes the modes show up better over small distances.

19.10.13 Pointing Instability Problems

Pointing instability shows up best with a quad cell alignment aid (see Section 12.9.11) and an oscilloscope or data logger.

19.10.14 Source Pulling

Etalon fringes can pull your laser frequency; of course, real Fabry–Perots are the worst for this (especially optical spectrum analyzers), but even stray fringes will do it. Faraday isolators are the best remedy if mild misalignment isn't enough.

19.10.15 Misalignment

We talked about alignment a lot in Section 12.4. One quick test for misalignment is to examine a focused spot on axis; a properly aligned optical system shows no asymmetric aberrations (e.g., coma) on axis and the beam doesn't move sideways when defocused.

19.10.16 Etalon Fringes

If you're suffering from etalon fringes in some component, then tapping that element should provoke a response, whereas components further downstream should be less sensitive (upstream ones may be as sensitive or even more, because moving them will move the fringe pattern). Going round tapping things gently with a screwdriver handle or the eraser end of a pencil is a surprisingly good diagnostic test. Etalon fringes are notoriously temperature sensitive, so gentle heating with your hand, a hair dryer, or a small heat gun (the little plastic kind you use for heat shrink tubing) will reveal a lot too. Etalon fringes forming across an air gap are sensitive to wind, temperature, and especially small amounts of vibration. With experience, you can locate the problem pretty rapidly this way.

Multiple effects cancel sometimes—a change that makes the fringes less vibration sensitive is probably an improvement, even if the fringes look somewhat worse.

19.10.17 Thermal Drift

There are lots of things that can drift with temperature besides etalon fringes; besides the obvious problems of expansion and bending, which we talk about in Section 20.2 (available at http://electrooptical.net/www/beos2e/thermal2.pdf), the two most drift-prone optical parameters are interference filter bandpasses and the focal lengths of infrared lenses.

19.10.18 Environmental Stuff

Dr. Erwin Loewen of Richardson Grating Laboratories used to tell a story of a famous ruling engine that would only work in the wintertime. It was a massive cast iron and granite apparatus, on the ground floor, and with a carefully temperature-controlled room.

After a few years of getting only 5 months' service per year, they finally nailed it by asking some very pointed questions about the day that the engine started misbehaving. It turned out that it usually happened several hours after the first big spring rainstorm, which pointed to something expanding due to moisture—but the iron, granite, and concrete seemed like poor candidates. Finally they took up the floor, only to discover that the engine had been mounted on several inches of wood over the concrete. The humidity changes made the wood swell and shrink, which was enough to prevent accurate gratings being made. Replacing the wood with something impervious fixed the problem.

19.10.19 Take It Apart and Put It Together Again

If all else fails, try taking the optical system apart, cleaning and inspecting everything, and putting it back together. Sometimes you'll find some hidden blunder, for example, a mount hanging from one overtightened screw and three loose ones, or a machining burr that didn't get removed. Other times, you won't know what changed, but the problem will vanish (those ones are usually the flaky ones due to etalon fringes or scattered light). Of course, sometimes the problem will get worse, and even if it does go away, you shouldn't ignore the problem—it's going to recur in another unit, count on it.