Chapter 2

DESIGN FOR MANUFACTURE

Barrie Gilbert
Analog Devices Inc.

2.1. Mass-Production of Microdevices

We generally think of mass production as a uniquely twentieth-century phenomenon. However, its evolution can be traced back much further. The explosion in printed books, following Johannes Gutenberg's fifteenth-century development of the Korean invention of movable type, had an impact on human society of heroic proportions. Precursors of modern mass-production, based on the *specialization of labour* and the *use of specialized machinery* to ensure a *high degree of uniformity,* can be traced to the eighteenth century. Writing in *The Wealth of Nations* in 1776, Adam Smith used the manufacture of pins to exemplify the improvement in productivity resulting from the utilization of uniform production techniques. Today, every conceivable sort of commodity is mass-produced. Pills, paints, pipes, plastics, packages, pamphlets and programs are mixed, extruded, poured, forged, rolled, stamped, molded, glued, printed, duplicated and dispatched worldwide on an immense daily scale. The most successful modern products are an amalgamation of many disciplines, years of experience, careful execution, rigorous production control and never-ending refinement.

In no other industry is the cross-disciplinary matrix so tightly woven, and the number of interacting elements so incredibly high, as in the semiconductor business. Reaching back to Gutenberg, and drawing on the principles of photography pioneered by Daguerre in the 1830s (embracing optics, lens-making, photosensitive films and chemistry), transistors are defined by a process of lithography, which is essentially *printing*. But what eloquent printing this is! A 200-mm silicon wafer has a useful area of about 200-cm^2, a little less than a page of this book containing some 400 words of text, equivalent to perhaps 16,000 bits. However, when divided into 1-cm^2 chips - the size of a modest microprocessor, today containing about 50 million transistors, through perhaps 20 successive layers of printing and processing - each wafer generates some 10 billion devices in a single mass-produced entity. In a production lot containing 40 such wafers, some 400 billion tiny objects are manufactured in a single batch. Multiply this by the daily manufacture of integrated circuits worldwide, and it will be apparent that the number of transistors that have been produced

C. Toumazou et al. (eds), Trade-Offs in Analog Circuit Design: The Designer's Companion, 7–74.

since the planar process was invented[1] runs to astronomical proportions far exceeding the expectations of its most optimistic and visionary progenitors.

Indeed, it is hard to identify any other mass-produced object that is fabricated in such prodigious quantities as the transistor. Even pills are not turned out in such numbers, and even when molecularly sophisticated, a pill remains a primitive amorphous lump of material. A transistor has a *complex fine-scale structure,* having a distinctive personality of its own (and a devious one: try modeling an MOS transistor!). Its near-perfect crystalline structure at the atomic level, and its precise dimensions and detailed organization at the submicron level, are fundamental to its basic function. No less important is the way these cantankerous virus-scale devices are tamed, teamed up and harnessed, in the *design ofmicro-electronic circuits.*

As their designers, we are faced with exciting opportunities and challenges. It is our privilege to turn essentially identical slabs of silvery-grey silicon – the stuff of mountains and the earth's most plentiful solid element – into clever, highly specialized components of crucial importance to modern life, handling everything from deceptively simple signals (voltages and currents, time intervals and frequencies) in analog ICs, all the way up to sophisticated packets of mega-information in computers and communication systems. Each of our creations will elicit uniquely different behaviour from the same starting material, and possess a distinctive personality of its own. How we shape this little piece of silicon, and the assurance with which it goes forth into the world and achieves its diverse functions, is entirely in our hands.

Integrated circuit designers who experience the rigour of dispatching their products to manufacturing, and watch them flourish in the marketplace and subsequently generate significant revenues for their company, soon discover that their craft entails a balanced blend of technique and judgment, science and economics. The path from concept to customer is strewn with numerous pitfalls, and it is all too easy to take a misstep. The practicing designer quickly becomes aware that silicon transistors, and other semiconductor devices, have a mind of their own, demanding full mastery of the medium if one is to avoid falling into these traps. One also learns that a circuit solution, no matter how original, elegant or intriguing, is of little value in abstraction. Cells, which will here be defined as small, essentially analog circuits of up to a dozen or so transistors, are merely *a resource* to be created (or discovered and understood), then tamed, refined and cataloged. Artful cell development is of fundamental importance to robustness in manufacture, but cells are certainly not the proper starting point for a *product* development, whose genesis arises within the context of broad commercial objectives, and which will exploit cell properties selectively

[1] By Jean Hoerni of Fairchild, U.S. Patent 3,025,589, filed May 1, 1959 and issued March 20, 1962.

and judiciously as the need arises. These basic fragments cannot be given any freedom to misbehave, if the products within which they are later utilized are to be manufacturable with high yields and at low cost.

This book is about how to design these basic cells so as to elicit some optimum level of performance, and particularly by considering the many trade-offs that invariably arise in adapting them to a specific use in a product. Such trade-offs are inevitable. Performance is always a compromise reached by giving up certain less desirable aspects of behavior in favor of those other objectives that are identified as essential. When such optimization is pursued with a set of public standards in mind (such as a cellular phone system like GSM), it is exceedingly important to find and utilize the "right" trade-offs, to provide an efficient and competitive design. Where the product is in the nature of a proprietary standard part, the choice of trade-offs may be harder, and involve more judgment and risk, since one often has considerable freedom to improve certain aspects of performance at the expense of others, in pursuing a particular competitive edge, which may be more sensed than certain.

For example, to halve the input-referred *voltage* noise spectral density in a bibolar junction transistor (BJT) low noise amplifier (LNA) one must at least quadruple the bias current.[2] However, this would be of little benefit in a cell phone, where battery power is severely limited, and provided that a certain acceptable noise figure is achieved, further reduction would be surplus to the system requirements. On the other hand, the same benefit would be very attractive in a state-of-the-art *standard product:* it could be the one thing that distinguishes it from all other competing parts. But then, with this increase of bias, the current-noise at the input port will double and that would no longer represent an optimal solution when the source impedance is high. While this is a rudimentary example of the pervasive "noise-versus-power" trade-off, decisions of this kind in the real world are invariably *multi-dimensional:* many different benefits and compromises must be balanced concurrently for the overall performance to be optimized for a certain purpose. It follows that trade-offs cannot be made in abstraction, in absolute terms; they only have relevance within the scope of a specific application.

2.1.1. Present Objectives

This chapter strives to illuminate the path to production a little more clearly, by providing a framework for successful commercial design. While it includes

[2] Specifically, the base–emitter voltage noise spectral density for a BJT due to shot noise mechanisms evaluates to $0.46\,\mathrm{nV}/\sqrt{\mathrm{Hz}}$ at a collector current I_C of 1 mA, and varies as $1/\sqrt{I_C}$. The current noise at this port, on the other hand, varies as $\sqrt{I_C}$. To these noise components must be added the Johnson noise due to the junction resistances, which does not depend to any appreciable extent on the bias current.

a few illustrative trade-offs, its emphasis on sounding down some more general tenets of robustness in cell design, with high-volume production in mind. The examples are drawn mostly from BJT practice. It outlines some basic cautions we need to observe in our design discipline, including our awareness of the limitations of device models and simulation, and examines the notion of worst-case design. Later, it delineates a dozen work habits of the manufacturing-oriented designer. A brief discussion of some of the ways we can minimize risk and optimize performance through the use of careful layout practices can be found in Chapter 33.

To reach the point of being ready to mass-produce a robust, cost-effective, highly competitive product, we will use many tools along the way. The best tool we will ever have, of course, is the magnificent three-pound parallel processor we carry on our shoulders. Nevertheless, for the modern designer, a circuit simulator, such as SPICE, when used creatively and with due care, can provide deep insights. Many brave attempts, including those of the author in his younger years, have been made to capture design expertise, in the form of programs that automate the design process. These range from such simple matters as calculating component values for a fixed circuit structure, to choosing or growing topologies and providing various kinds of optimization capabilities. Advanced design automation works well in coping with procedures based on clearly-defined algorithms, of the sort that are routine in digital design. However, they have been less successful in aiding analog design, and are of little help in making trade-offs. This is largely because each new analog IC development poses distinctly different design challenges, often calling for on-the-spot invention, since cell reutilization is fraught with problems and of limited value. In this field, as elsewhere, there are no algorithms for success: we must continue to rely on our creativity, our experience, our ability to draw on resources, and our judgment in facing the matter of design trade-offs.

Numerous pitfalls and obstacles will be encountered on the path between the bright promise of the product concept and that moment the IC designer most looks forward to: the arrival of first silicon. But the seasoned engineer knows that these first samples are just the tokens we handle at the beginning of a longer and more arduous journey. Still ahead lie many months of further documentation and extensive testing, during which the glow of early success may fade, as one after another of the specifications is found to be only partially met, as ESD ratings are discovered to be lower than needed on some of the pins, or as shadowy, anomalous modes of operation make unwelcome cameo appearances. There follows the challenge of finding ways to make only minor mask changes to overcome major performance shortfalls; the interminable delays in life test; and the placating of impatient customers, not to mention the marketing folks, who see the window of opportunity at risk of closing.

2.2. Unique Challenges of Analog Design

Such obstacles stand in the way of all professional IC designers, but there are radical differences in individual design style, and between one sub-discipline and another. In the digital domain, the design focuses on assembling many large, pre-characterized blocks, comprising thousands of gates, amounting in all to a huge number of transistors (often known only approximately[3]) each one of which must reliably change state when a certain threshold is reached. Advances in this domain stem largely from improvements in micro-architecture, a relentless reduction in feature size and delay times, and advances in multi-layer metalization techniques, which are also necessary to pack more and more functional blocks into the overall structure, while keeping the chip size and power to manageable levels.

As clock rates climb inexorably into the gigahertz range, the dynamics of these gates at the local level, and the communication of information across the chip, are generating problems that, not surprisingly, are reminiscent of those encountered in classical RF and microwave design. Further, the very high packing densities that are enabled by scaling give rise to new problems in removing the heat load, which, milliwatt by milliwatt, adds up to levels that demand special packaging and sophisticated cooling techniques. Such issues, and the sheer complexity of modern microprocessors and DSP elements, will continue to challenge digital designers well into the century. Their trade-offs will not be addressed here.

The challenges that arise in the domain of analog functions are of a distinctly different kind, and stem principally from two unique aspects of analog circuits. First, there is *much greater variety,* both in chip function, which can take on hundreds of forms, and in the particular set of performance objectives, and even the specification methodology (such as "op-amp" versus "RF" terminology), from one product to another. Second, the actual performance, in all its many overlapping and conflicting facets, *depends on the detailed electrical parameters ofevery one ofthe many devices* comprising the complete product, and in a crucial way for a significant fraction of this total. Obviously, it is quite insufficient to simply ensure that a transistor is switched on or off, or even that this transition occurs very quickly and at just the right time; such are only the bare bones requirement of the analog transistor. So much more is now involved in "meeting the specs", and *thisparametric sensitivity* touches at the very heart of what makes analog circuits so different from their distant digital cousins.

[3] Patrick Gelsinger of Intel told me the exact number of transistors in the 486 micropocessor is 1,182,486 (the last three digits were "a coincidence") noting that how one counts devices is somewhat imprecise in the first place.

Much of what we do as designers will require constant vigilance in minimizing these fundamental sensitivities.

Many detailed challenges in signal management face the analog designer. In even a simple cell such as an amplifier, one is confronted with first, the choice of a topology that is both appropriate and robust; then the minimization of noise, distortion, and power consumption; maintenance of accurate gain; elimination of offsets; suppression of spurious responses; decoupling from signals in other sections performing quite different functions; coping with substrate effects; unrelenting attention to production spreads, temperature stability; the minimization of supply sensitivity, and much more.

In the domain of nonlinear analog circuits, special effort is needed to achieve accurate conformance to one or more algebraic functions, such as square-law, product and quotient, logarithmic and exponential responses, and the like. With all nonlinear functions there is also a special need for vigilance in the matter of scaling, that is, control of the coefficients of the contributing terms. Voltage references are often needed, which may need to be exact without recourse to trimming. In filter design, another set of imperatives arises, having to do with ensuring accurate placement of the poles and zeroes of the transfer function even in the presence of large production tolerances. Many modern products combine several of these various functions, and others, in a single chip.

Hard-won analog design victories are known only to a small group of insiders, who are proudly aware of the continual, quiet improvements that so often are behind many of the more visible successes that shape modern communications devices, and which are likely to be bundled with the DSP and microprocessor parts of the system and presented to the public in the guise of yet another advance arising solely from the wondrous properties of digital technologies. One can understand the indifference to analog techniques invariably displayed by the public, but it is worrisome to see this now appearing in the attitudes and skill-sets of new graduates in electronics. Behind all of the glamor that digital systems generate in the popular eye, there is a massive infrastructure of essential analog electronics, and a growing need for skilled analog designers. In the twenty-first century, design challenges with a pure-analog emphasis will not diminish; rather, they will be plentiful. Unfortunately, the number of new engineers available to address these challenges may not keep up with the demand. University students are often led to believe – incorrectly, just like the public at large – the now familiar mantra that “analog is obsolete.” This is manifestly false.

These challenges will continue to be related to achieving *small but exceedingly difficult improvements* in certain key parameters, rather than increasing the raw number of transistors that can be crammed into the latest CPU. For example, while a 1-dB improvement in the signal-to-noise ratio of a receiver does not seem very impressive, it typically results in a ten-fold improvement in the bit-error-rate of a digital channel. It requires considerable inside knowledge to separate the confusing claims made for the latest digital gadget, so persistently and persuasively made by their promoters, from the fact that analog techniques remain important even in the most sophisticated of these products.

The common view is that, by virtue of the certainty of binary data, digital systems avoid the many ambiguities of analog circuits, which have a reputation for being unrepeatable, temperamental, unstable, prone to drift and loss of calibration, or bursting into oscillation without warning. Many of these weaknesses are real, and can be traced to poor design, particularly through inattention to the all-important matter of *robustness* and the *minimization of parametric sensitivities,* which is why there is a need for a book of this sort. Nevertheless, a crucial dependence on the precise values of certain dimensional parameters – for example, those determining the bandwidth of an amplifier – is frequently unavoidable, and unrelenting vigilance is needed during design to ensure robustness in production. Close attention to component tolerances and design margins is essential, and trade-offs must be made carefully.

For example, it is soon discovered that there are inherent trade-offs to be made between achieving uncompromising state-of-art performance on the one hand, while minimizing cost and ensuring a high degree of robustness and chip yield on the other. Since this is true, modern system designers are only being prudent in seeking ways to reduce the "analog front end" to the barest minimum, or even eliminate it; invariably, they are not being unfair in asserting that "This is where our worst problems are to be found." Analog circuits will always be prone to these criticisms, because they are fundamentally closer to the physical reality than are digital circuits. And this is where another key difference is to be found.

2.2.1. Analog is Newtonian

In an important sense, *analog circuits are closer to nature than are digital circuits.* This viewpoint can help us to understand why these two domains of endeavor are fundamentally so different.[4] Certainly, many of the challenges in digital electronics today also have a strongly physical aspect, mostly, although not entirely, at the cell level. But these stand apart from the more important development thrusts relating to the transformation of logical data, rippling through gates which reshape and retime this data, within which the strictures of sequential discrete algorithms replace the unfettered autonomy of the analog

[4] There are actually *three* fields of electronics today: the two major groupings, analog and digital, and a third, smaller but well-defined and rapidly-growing group of techniques which we can call *quasi-analog* or *binary-analog,* exemplified by "sigma–delta" techniques. The three basic disciplines overlap strongly and are co-dependent: they are at once symbiotic and synergistic.

circuit. Once a library of digital cells has been generated, with careful attention to time delays and threshold margins, their inherently analog nature is no longer of interest in digital design.

Analog circuits are more deeply allied to the physical world because they are concerned with the manipulation *ofcontinuous-time, continuous-amplitude* signals, often *ofhigh accuracy,* having *dimensional* attributes, traceable *tofundamentalphysical constants.* (Logic signals are, of course, dimensionless.) The primary physical units are length [L] in meters, mass [M] in kilograms, and time [T] in seconds, and we here use charge [Q] in coulombs as the fourth basic unit.[5] The *physical algebra* of analog-circuit analysis differs from ordinary algebra in requiring attention to *dimensional homogeneity.* Thus, *voltage* signals embed the dimensions of $[ML^2T^{-2}Q^{-1}]$. Sometimes, greater importance is attached to the signal *currents,* which are of dimension $[QT^{-1}]$. Voltages are just another way of representing *energy* $[ML^2T^{-2}]$ normalized through division by the electron *charge* while current may be envisaged as counting multiples of charge quanta over a specified *time* interval. It follows that *current-mode* signal representation is more prone to absolute-magnitude errors than voltage-mode representation, since in the latter case, scaling can be quite directly traced to such things as the bandgap energy of silicon, the Boltzmann constant *k,* temperature and electronic charge, *q.* Nevertheless, current signals can maintain high *ratio accuracy* and have certain benefits.

Dimensional quantities are inextricably woven into the fabric of the universe, from sub-atomic forces up to the largest cosmic objects. They are also embedded in energy fields. RF signal levels in a transceiver can be equated to an electromagnetic *field strength* at the antenna, and expressed as a *power,* $[ML^2T^{-3}]$, at *some frequency* $[T^{-1}]$. Similarly, the electrical circuit elements within which these signals flourish and propagate have their own set of physical dimensions: *resistance* $[ML^2T^{-1}Q^{-2}]$; *capacitance* $[M^{-1}L^{-2}T^2Q^2]$; and *inductance* $[ML^2Q^{-2}]$. The attribute of *spin,* $[MLT^{-1}]$, is an essential aspect of semiconductor device behavior, as are the *mass* [M] and *velocity* $[LT^{-1}]$ of holes and electrons, and the pure *length, width and thickness* [L] of device structures. In view of this strongly-physical nature of analog circuits, it is not inappropriate to use the term *Newtonian* to describe them.

2.3. Designing with Manufacture in Mind

Designing integrated circuits in a commercial context, one is daily confronted with the need for compromise, expediency and pragmatism – which

[5] The International System of Units (SI) uses the Ampére, rather than charge. Charge is used in the present context because it is an intimate aspect of semiconductor physics.

continually orbit our concerns about development time and product cost – while preserving performance and robustness. These imperatives are rarely addressed in technical university courses. It is common to pursue only those aspects of design which one most enjoys, such as exploiting an exotic new technology, conceptualizing intriguing and bold new approaches, constructing grand system architectures, devising new circuit functions, discovering novel topologies, laying down a fine theory, acquiring a patent or two, or writing a paper for a major conference or professional journal. At times one may lean toward a highly favorable, idealized viewpoint of the task, deferring criticism and "second order effects" for another time. If not careful, one may completely lose sight of the fact that the variables which are so confidently manipulated in spread-sheets and simulations (gain, noise, intermodulation, power, matching and stability criteria, bandwidth, phase margin, frequency, and the like) are but a simplification of harsher realities.

Assailed by all the slings and arrows of outrageous wafer processing, products conceived in the refined conceptual world face a traumatic trial, which only the fittest survive. While intellectually aware that this is so, we may pursue our design work with optimism, in the tacit belief that our devices are basically uniform and predictable, and element variability is only a secondary consideration. Because of the tight controls on the many steps used in a modern IC process, this is not an entirely vain hope. We have come to expect extraordinarily high manufacturing standards and prodigious production yields, often to exacting specifications. Nevertheless, many disappointments can creep into the performance of production components. Some of these are certain but unavoidable; others, while equally predictable, can be averted by the use of thoughtful design practices. Often, we have to sacrifice certain desirable aspects of performance to ensure some others will be met, the essence of a trade-off, which is the central theme of this book.

2.3.1. Conflicts and Compromises

In the world of commercial product design, *performance trade-offs are rarely two-fold in nature.* Certain design conflicts arise in pairs when utilizing a given technology, such as between bandwidth and power consumption, between intermodulation and noise, in balancing the contributions of voltage and current noise, and so on. But these can just as easily be coupled in other ways: noise is in a constant contest with bandwidth; intermodulation distortion can often be lowered only by using higher power consumption; and many aspects of static accuracy are in conflict with achieving high bandwidths. Each design involves complex, multi-variable interactions, and compromises are inevitable.

Good practice demands that adequate consideration is given to every one, perhaps hundreds of such conflicts that can arise during several weeks of design time, sometimes within the compass of a dozen transistors. Indeed, as we shall see later in this chapter, even a one-transistor LNA can consume a great deal of effort in order to optimize its performance and to be able to guarantee that it will fully meet all of its specifications in every one of millions of future instantiations of the product in which it is embedded.

A thorough understanding of these interactions is the essential starting point in the long road to design mastery in the analog domain. A very basic consideration is that of suppressing, as far as possible, the effects of temperature on circuit behavior. The second most obvious objective is to minimize the impact of changes in supply voltage. And even when suitable countermeasures have been found, and all the fundamental circuit relationships have been aligned in the most optimal manner for a particular set of objectives, there remains the significant hurdle of desensitizing performance to *production variances.*

These three top-level obstacles to achieving robust and reliable performance are sometimes referred to as the PTV (Process, Temperature, Voltage) aspect of the design challenge. Beyond these barest of necessities lie the broad plains of optimization, the central design phase in which performance conflicts will met by making trade-offs. However, before we can proceed with a detailed discussion of some examples, and start to think seriously about optimization, we must give further consideration to the various types of process sensitivities that can arise in analog design. Further, it must be understood that these are in no sense sequential parts of a design flow, during which each potential sensitivity, or an aspect of optimization, is addressed and then set aside. Undesirable circuit interactions can appear at any time. The most dangerous are those which arise due to "trivial" changes made late in the design process, changes that are in the nature of an afterthought, and which thus do not receive the benefit of the thousands of hours of simulation studies that probably went into shaping the rest of the product, and rigorously verifying its behavior.

2.3.2. Coping with Sensitivities: DAPs, TAPs and STMs

In a typical IC manufacturing process, there are numerous production parameters that vary, including: implant dose rate and time, and other factors affecting total doping concentrations; furnace temperature and time; gas flow rates; etch and deposition times; resist composition, and other factors related to chemical quality; oxide growth rates, fine structure and uniformity; resist thickness and uniformity; micro-assay composition of sputtering targets; and so on. These "low-level" physical variations will manifest themselves through an even wider variety of effects in the "high-level" electronic parameters at the device level.

Beyond this, the use of numerous different circuit topologies in the design phase, and the broad and essentially unconstrained choice of operating conditions for each device, create even greater *parametric complexity.* It is inevitable that these variances will influence the "top-level" performance of our circuit, to a greater or lesser degree. We have to allow these variances full reign, while ensuring that nearly every instantiation of the product across the wafer meets its operational specifications (which is the first aspect of the *robustness* challenge) and that every sample passing muster during production testing will remain within its performance limits over its lifetime, when large temperature and supply voltage variations can occur (the second aspect of the robustness challenge).

Success in this context requires attention to the most minute detail, and may easily fall out of our grasp, if even a seemingly minor detail is neglected. The simplest of components, such as monolithic resistors and capacitors, embody numerous low-level process parameters which influence their absolute value. Suppose that we are relying on a resistor–capacitor product to determine a time-constant, and thus set the frequency of an oscillation. We must design our product so that the error in the unadjusted frequency can be accommodated; that is, either we can formulate a method for manually trimming to the needed accuracy, or the worst-case[6] uncertainty is within the capture range of some automatic tuning means. Errors in the resistor and capacitor contribute equally to the error in frequency, which is of the form k/CR.

Most basically, the sheet resistance of the layer used for fabricating the resistor is subject to considerable variation. In a diffused or polysilicon resistor this will arise from variations in doping concentration and the depth of the diffusion or film, and can easily be as high as ±15%, a 30% spread. Conductance in any resistive layer is also a function of temperature, sometimes a strong function. For example, the sheet resistance of a diffused resistor may typically vary by 1,500 ppm/K at $T = 300$ K, which extrapolates to variation of about 20% over the 130 K range from 230 to 360 K (–43°C to 87°C). This raises the tolerance band to about 50%. Hopes of containing the frequency within a narrow range are already fading.

Variations in the width and length of the resistor must also be accommodated. When the absolute value needs to be well controlled, one would normally choose to use a physically large resistor, but this may be contraindicated when operation at high frequencies is also required, and the parasitic capacitances of the resistive layer become prohibitive. Assuming a moderate width of about $10\,\mu\mathrm{m}$ for such a situation, and allowing for a maximum variation of $0.25\,\mu\mathrm{m}$ at each edge, we are faced with a further 5% uncertainty. There may also be some voltage modulation of resistance. Thus, the resistance alone may vary over a

[6] The question of whether the term *worst-case* always has a definite meaning is discussed later in the chapter.

60% range, in a high-volume, robust design context. Adding to this estimate all the similar variations in the capacitor value, particularly those due to variations in the dielectric layer, and for junction and MOS capacitors their varactor behavior, it is easy to understand why the frequency of our basic oscillator can be predicted in only approximate terms: it already has process, temperature and possibly supply sensitivities even before considering the effect of the active elements. Specifications based on the assumption of tighter controls are worthless.

This is a very common situation in analog design, and stems directly from the physical nature of analog signals and components. Aspects of performance that exhibit this particular kind of sensitivity can be classified as *Dependent on Absolute Parameters;* we will refer to such aspects of performance as "DAPs". It is impossible to eliminate sensitivity to this class of parameters by design tricks, though we may in special cases be able to reduce the sensitivity.

For example, the gain–bandwidth of an IC operational amplifier invariably can be traced to the product of a resistance (ultimately setting the value of a g_m) and a capacitance (which may be defined by an oxide layer, as would usually be true for a low-frequency op-amp, or an incidental junction capacitance, as might be the case for a wideband amplifier). Since even carefully designed resistors may have a tolerance of up to ±25%, and capacitors can vary by ±15%, the control of gain–bandwidth in an op-amp[7] may be no better than ±40%. However, it is later shown that when using this amplifier cell in a closed-loop mode, one can introduce a lag network into the feedback path such as to implement an overall two-pole response just above the high-frequency roll-off in which the gain at some (known) signal frequency can be made much less dependent on the position of the dominant pole. The method invokes the reliable *matching* of similarly-formed components, the cornerstone of all monolithic design, to lower the sensitivity to their actual values, in a rather non-obvious way.

In the fastest amplifiers we can make, using BJT processes, and in which the transistors are operating near their peak f_T, it is more likely that the variations in effective base-width and current density cause the production spreads in bandwidth. In turn, the current density depends on the actual emitter area (thus, on lithography) and is invariably dependent on some on-chip voltage source and at least one resistor. Since the f_T is a diminishing function of temperature,

[7] Few op-amp data sheets are forthcoming about this spread, often stating only a typical value. Similar vagueness is often found in the specifications for RF products. Some of this imprecision can be traced to the cost of testing ICs to allow these aspects of performance to be fully guaranteed; some of it has arisen as a kind of tradition, with concerns that the explicit revelation of the magnitude of such spreads would put a more completely-specified part in a "bad light".

spreads from this source must also be addressed. In those cases where devices are operated at very low currents, however, the device's g_m, its (uncertain and voltage-dependent) junction capacitances, and interconnect capacitances set a limit to attainable bandwidth. Whatever the precise mechanisms, the bandwidth of virtually all monolithic amplifiers is strongly "DAP", and in system design we must find ways to accurately define the channel bandwidth (which is only a fraction of the amplifier bandwidth) by the use of off-chip components, such as LC resonators, SAW or ceramic filters, or high-precision CR networks. Certainly, it would be very unwise to depend to any critical extent on the unity-gain frequency of common feedback amplifiers.[8]

As a rule, most (though not all) specifications which have a dimension[9] other than zero will be DAPs. These include time $[\mathrm{T}]^1$ and frequency $[\mathrm{T}]^{-1}$; current $[\mathrm{A}]^1$ in a cell (setting g_m and total consumption); all internally-generated voltages $[\mathrm{V}]^1$, (such as noise, V_{BE}, bandgap references, etc.); inductance $[\mathrm{L}]^1$; capacitance $[\mathrm{C}]^1$; resistance and impedance $[\Omega]^1$ or $[\mathrm{L}]^{0.5}[\mathrm{C}]^{-0.5}$; conductance and admittance $[\Omega]^{-1}$ or $[\mathrm{L}]^{-0.5}[\mathrm{C}]^{0.5}$; etc. These sensitivities are addressed in various ways, some well known. Where absolute accuracy is essential, we can bring the dimension of "time" to an IC by utilizing a reference frequency defined by a crystal; or we can introduce the International Volt by laser-trimming against a primary standard during manufacture; we can use external resistors to establish accurate currents; and so on.

Next we turn to the second of these sensitivities. Absolute errors in the element values of *all components made of the same materials* (of all resistors, all capacitors, all current-gains, all V_{BE}s, etc.) need not affect certain crucial aspects of performance. By relying on the use *of pure ratios,* we can assure the accuracy of any specification having dimension zero. Examples are gain at relatively low frequencies (and gain matching); attenuation (even up to high frequencies); relative phase between two signals (and precision in quadrature); filter Qs and overall filter shapes; conformance to functional laws (such as logarithmic, hyperbolic tangent, square-law); waveform, duty-cycle, weighting coefficients; DAC/ADC linearity, and the like.[10]

[8] In the 1970s a great deal of nonsense was being published about using "the operational amplifier pole" as a basis for the frequency calibration of what were misleadingly called "Active-R Filters".

[9] Again, the dimensions used here are those familiar to electrical engineers. In a formal treatment, they would of course be expressed in fundamental MKS or CGS units. Logical signals have dimension zero.

[10] Of course, the use of digital ratios brings an even higher level of accuracy, for example, in frequency division. But not all logical circuits are above reproach. Phase jitter and non-quadrature are just two examples of error in supposedly pure-binary circuits where analog effects lead to degraded performance.

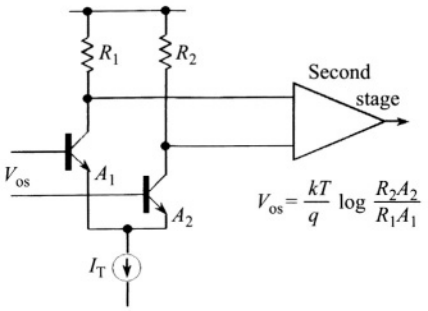

Figure 2.1. Some ratiometic circuits produce dimensional quantities.

We may call such specifications *Tolerant to Absolute Parameters,* and will refer to them as "TAPs". Because of this tolerance, or low sensitivity to tracking element values, we can in principle achieve highly accurate low-frequency gain, even in the presence of large absolute variations. In special cases, even some dimensional variables are in this class of TAPs. For example, the input-offset voltage of an op-amp using a BJT differential-pair as its g_m stage (Figure 2.1) is a precise function of the circuit parameters:

$$V_{os} = \frac{kT}{q} \log \frac{I_1 A_2}{I_2 A_1} = \frac{kT}{q} \log \frac{R_2 A_2}{R_1 A_1} \tag{2.1}$$

Provided that the emitter areas and the load resistors can each be made closely equal,[11] the offset voltage will be small, typically sub-millivolt. Its actual magnitude will be dependent on neither the absolute size of the emitters nor the absolute value of the resistors, and it is scaled only by the fundamental dimensional quantity *kT/q* (25.85mV at $T = 300\,\text{K}$). In monolithic analog design, we are constantly on the lookout for phenomena of this sort. The TAP perspective relies on a strong reliance on *ratios* to eliminate the effect of large absolute variations in parameters, and on an appeal to fundamental scaling phenomena rather than a reliance on external stimuli.

A related use of the above equation is the generation of a bias voltage based on the "delta-V_{BE}" idea, in which the emitter-area ratio A_2/A_1 (sometimes in combination with the resistor ratio R_2/R_1) is deliberately made much greater than unity. For example, when the net ratio is set to 48, V_{os} has a theoretical value of 100.07 mV at 300 K. This voltage will be proportional to absolute temperature (PTAT), which is often the most suitable biasing choice in BJT design. Since its basic value can be precisely determined by a pure ratio, and subsequently

[11] This is a simplification; other factors, including base-width modulation (early voltage) and various on-chip gradients are involved.

multiplied up to a higher value, better suited for IC purposes (say from 300 mV to 1 V) by another pure ratio, we can fundamentally eliminate the sensitivity to absolutes. Incidentally, it will be apparent that the delta-V_{BE} concept can be used as the basis of a silicon thermometer, and when implemented using more careful techniques than briefly described here, the voltage can be accurate to within 0.15%, corresponding to a temperature error of < 0.5 K at 300 K.

Finally, in this set of process sensitivities, we must address aspects of circuit performance that are *Sensitive To Mismatches,* which we call "STMs". Clearly, this includes a great many effects, since the immunity conferred on a circuit function through the use of pure ratios is immediately lost if these ratios are degraded by mismatches. (As used in this frame of reference, the term refers not only to components that should be *equal,* but to the deviation from some *nominal ratio.*) Here again, the strongly Newtonian nature of analog circuits is apparent, since matching accuracy is directly related to device size. It is clear that the greater the number of atoms used to define some parameter, the lower the sensitivity to absolute variations in this number. We are here faced with a very basic trade-off, since the use of large devices, whether passive or active, is at odds with the minimization of inertia,[12] and also with the minimization of die size. In fine-line processes, one is inclined to use small geometries rather uniformly, to achieve the highest speed and packing density; but high accuracy analog design requires careful attention to the *optimal scaling* of devices. Bigger is not necessarily better, however. Even when die size and device parasitics are not critical considerations, the use of excessively large devices can actually cause a *reduction* in matching accuracy as various gradients (doping, stress, temperature, etc.) begin to assert an influence.

This interdependence of circuit design and layout design is found in all integrated circuit development, and serious lapses will occur if they are ever treated as separate and distinct activities, but especially so in analog design. There are many times when one can achieve a very distinct advantage, whether in speed, accuracy, packing density, or robustness, by altering the circuit design to accommodate a more promising layout scheme. Further, the generous use of similar device orientations, sets of physically parallel resistors, and dummy components at boundaries pay significant dividends in preserving analog accuracy.

With some thought, it may be possible to actually avoid the need for transistor matching at all, through the use of dynamic element matching, based either on the better matching that can be achieved between capacitors, or through the use of clever switching of the topology, either to alternate error sources in a

[12] A general term favored by the author to describe the net effect of all mechanisms leading to the storage of charge in a device, which causes sluggishness in the response.

canceling fashion, or by an appeal to averaging. Thus, the most accurate silicon thermometers do not depend on the (still somewhat risky) matching between two separate transistors, which can also be degraded by mechanical strain across the die. (Transistors are always willing to operate as strain gauges.) Instead, a single junction can be used, and biased sequentially at two or more current levels. The integer ratios between these excitation phases can be generated to very high accuracy. The resulting small PTAT voltages are amplified, and subsequently demodulated, by switched-capacitor techniques.

One can implement dynamic band-gap references using similar methods, although in this case there remains an unavoidable dependence on the actual saturation current of the junction, $I_S(T)$, which is always a matter of total doping level and the delineation of the junction area. While this DAP remains, there are still further tricks up the analog designer's sleeve to reduce these sensitivities in the design of advanced band-gap references, from "direct" to "diluted", but they cannot be fully eliminated. One can see why this is so, by remembering that the transistor is used essentially as a *transducer,* from the domain of temperature to the domain of voltage. Also, since this $V_{BE}(I, T)$ is dependent on the *absolute current density* in the device, which in turn depends on some on-chip resistor, it can be stated with certainty that there is no way to design a reference to be inherently traceable to a fundamental physical constant such as the bandgap energy of silicon.

In making trade-offs in device structure, scaling and placement for analog design, one can appeal to *principles* and *guidelines,* but it is unwise to rely on *rules.* Some of the principles of matching are obvious and unequivocal; others tend to be wrapped in folklore, a reflection of the common fact that insufficient statistical data is available to state much with certainty, in many practical cases. This is often because one is designing on a new IC process for which statistically-reliable data has not yet accumulated. Guidelines for matching, which is not a matter of basic circuit design but rather, the design of the layout, are provided in Chapter 33. However, absolute attention to device sizing must be made during the design phase, and very definite parameters assigned to all components prior the Design Review, since one cannot assume the layout designer is a mind-reader. These should not only be embedded in electronic form, in the captured schematics, but should also be immediately visible on these schematic, in the pursuit of total clarity and the elimination of ambiguity, as well as in the spirit of full disclosure of all design issues for peer review, and possible correction.

2.4. Robustness, Optimization and Trade-Offs

The expression *robust design* is widely used. We have an intuitive sense of what this means and entails. A robust product is one whose design ensures that it is not critically dependent on the precise materials used in its construction, and is able to fully perform its intended function under all anticipated operating conditions and endure vigorous environmental forces without significantly affecting its long-term utility. In civil engineering, such as the construction of a major bridge, these would include a consideration of material stress limits in the presence of worst-case traffic loading or unusually severe cross-winds, recognizing the criticality of choosing the construction materials and the actual process of fabrication.

The trade-offs related to robustness that go into the design of a modern ICs are at least as numerous as for large engineering projects, such as bridges and buildings. They may also involve similar concerns for product liability, for example, in components used in medical equipment, or where electromagnetic emanations may pose a threat to a human user.

A robust circuit design is one in which the sensitivities of critical performance specifications to variances in the manufacturing process and the circuit's operating environment are first fully anticipated and identified and then systematically nulled, or at least minimized, through optimal choices of macro-structure, cell topology, individual device design, component values, bias conditions and layout.

Can we define a "Robustness Coefficient"? Almost certainly not. Even some sort of "Figure of Merit" is unlikely. Can we delegate the maximization of robustness and its inverse, the minimization of sensitivity, to a computer? Only in a few special and limited situations. This is where one's mastery of design will play its most indispensable role. Time and again, we find that the search for the most robust solution requires that we know how to shift attention, as circumstances require, from the *whole* to *the parts* and back again to the *whole* – numerous times in the course of the product development. There is a fractal-like quality to analog IC design, in the sense that whether we are viewing it at a high level, wearing the customer's shoes, or stepping down through many layers of circuit structure and operation, the biasing of its components, device optimization, the physics at the next layer below that, there is at every level a huge amount of information to consider and a great deal of complexity to cope with.[13]

It is important to understand the distinctions between robustness, optimization and trade-offs. While these topics overlap very considerably, they stem

[13] Again, we may note that, once one gets down to the gate level, there is little to be gained, in the pursuit of digital system design, by probing deeper into structure.

from quite different impulses. As we have seen, robustness is a *state*; it is the outcome of pursuing analytical methods, simulation studies, and the selection of technologies, architecture and scaling and judgments in the course of a product design. The threads leading to this result will lead back to many sources, but most notably from the pursuit of optimization and the making of trade-offs.

Optimization is a *process*. It is the analytical consideration of a system and its parameters with a view to discovering local minima and maxima in n-space (where in practice n is often much greater than 2) which can be identified in some particular way as the best choice(s), where the performance aspects of special interest are closest to what can ever be achieved within the constraints of a given architecture, technology or specific component limitations. This is a methodical, systematic process very amenable to mathematical representations or, more commonly, numerical methods. Thus, optimization is an *algorithmic process.* Since the representational equations "know" nothing about the world beyond their n dimensions, there is no expectation of discovering new worlds of possibility; maxima and minima never turn into wormholes.

Consequently, one can never be sure that the solution offered by an optimization process is truly the *best of all possible choices*: it is only the best of a severely limited sub-set of choices. In this sense, it is as much the product of the framer of the algorithm as of the data. Further, numerical optimization provides little if any insight into extending performance beyond these boundaries, and because an analysis does not include all the variables, it may not even be finding the actual best case in practice. This will frequently be true even for rudimentary circuits, such as a cell-phone power amplifier. Finally, there is a strong likelihood that the under-skilled user of optimization procedures ("design programs") will believe that the "answer" is genuine and reliable, while learning nothing in the process.

In contrast, the act of making a trade-off is *no sense* algorithmic. Trade-offs require a human *decision,* namely, the difficult and vexing choice between two or more *equally attractive alternatives,* and the sacrifice of one good for another. It is a zero-sum game. It involves risk and calls for judgment. In this common situation, there are no rules to lean on; if there were, the next step in a design would not be a trade-off, but the mechanical, unthinking application of some such rule. In the end, all decisions are emotional.[14] Many engineers are inclined to reject this tenet, proclaiming that this may be so in the social world, but not in technology, where each step in a development proceeds logically. However, it does not take many years practicing design to see the truth of this statement. When all the evidence, facts and analyses point clearly and unequivocally to a single, definite course of action. no decision is needed: that

[14] Due to Edward de Bono, a professor of psychology at Oxford University.

is optimization. But in the many cases where the data are flat, equally favouring many possible ways forward, a decision is called for. *That is* a trade-off.

It may even be a coin toss. In developing a standard linear product, having a wide applications domain, but lacking all the required market data, the designer is often forced to make guesses, based on personal "market savvy" and experience as to the most useful combination of performance parameters. One frequently needs to decide whether to pitch the product toward leading edge performance and stop worrying about its 50 mA supply current, or toward portable applications, by halving the current and accepting that performance will suffer. Similar trade-offs will arise between using bare-bones, ultra-cheap design practices with a view to achieving the smallest possible die area, in order to be competitive in pricing the product, or err on the side of extending the feature set and improving the performance, to extend the applications space, and considering such factors as ease of use and customer satisfaction. There are no algorithms for success.

2.4.1. Choice of Architecture

We will now look at several case histories, to illustrate the meaning of robustness in more concrete terms. In doing so, we will appreciate how elusive a quality it can be. To achieve the most satisfactory overall solution requires that numerous parallel and competing factors need to come into focus into a unified vision of the whole. Many trade-offs, which are open-ended decisions, are needed.

Clearly, we need to start with a robust architecture. Of the numerous ways we can satisfy a system requirement, some will be more sensitive to slight changes in parameter values than others. A simple example is provided by a cellular phone system involving a limiting IF with an received signal strength indication (RSSI) output (Figure 2.2). In this example, the RSSI output voltage – which reports to the cell supervisory system the strength of the received signal, in order to minimize the transmit power in the handset and at the base station – is scaled by a band-gap reference voltage, generated in the receiver sub-system. This voltage is then measured, and converted to digital form, by another IC, a codec, in which a second bandgap generator is embedded. Either or both of these circuits may be built in CMOS, a technology which is not noteworthy for high reference-voltage accuracy.[15] A guaranteed absolute accuracy of ±5% in

[15] See B. Gilbert, "Monolithic voltage and current references: theme and variations," in: J. H. Huijsing, R. J. van de Plassche, and W. M. C. Sansen (eds), *Analog Circuit Design,* pp. 269, which includes further examples of good and bad planning in the use of voltage references.

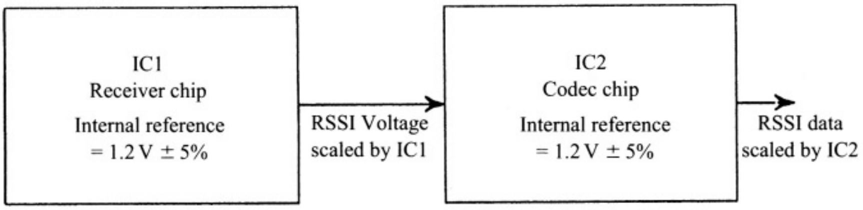

Figure 2.2. A vulnerable approach to scaling of nonlinear circuits.

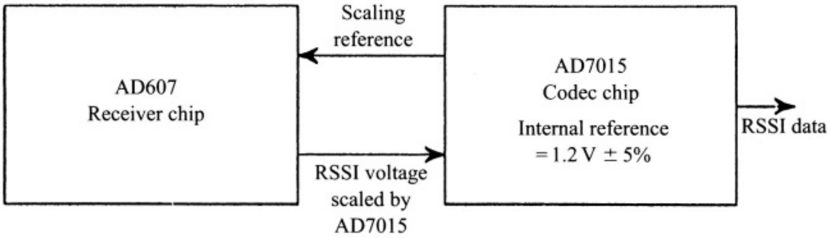

Figure 2.3. Use of a common reference voltage removes uncertainties in scaling.

each reference is a reasonable objective if high yields are to be achieved and the cost objectives do not allow trimming.

There could have been historical reasons for the use of this approach. For example, one circuit may have been designed ahead of the other, as part of a separate venture. Clearly, in this scenario, there is a worst-case error in the RSSI calibration of ±10%. If this occurs at the top end of a receiver's 70 dB dynamic range, the measurement error could amount to ±7 dB. In this scheme, there is also some yield loss due to the use of at least one redundant reference generator. Finally, it is possible that the uncorrelated noise of the two independent references could lead to LSB instabilities in the measurement; this may be especially troublesome where there is a high level of flicker noise, as in a pair of CMOS bandgap references.

Figure 2.3 shows a first alternative, in which only a single reference is used. This method is used in the Analog Devices AD607 single-chip superhet receiver. The mixer and linear IF strip are provided with a linear-in-dB gain control (AGC) function, the scaling reference for which is derived from the companion codec (AD7015). The error in that reference is now inconsequential, since it alters both the scaling of the RSSI output (so many mV/dB) and that of the ADC in the codec (so many LSBs per mV). Here, we have a classic example of the minimization of sensitivities through a dependence on ratios at the *system level.* The revised approach can allow much looser tolerances on

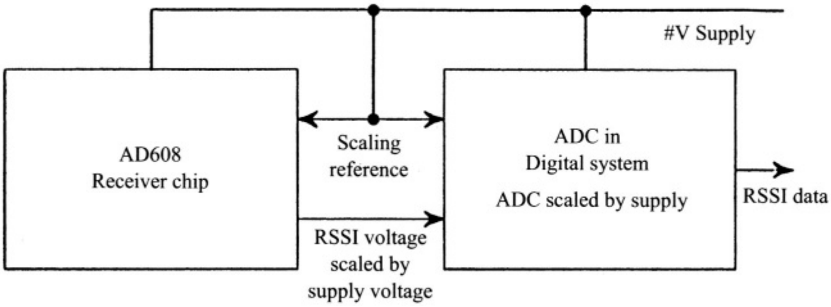

Figure 2.4. Absolutely calibrated voltage references are often quite unnecessary.

the remaining reference, if accuracy is not needed for any other purpose. Close matching of resistor ratios (utilizing unit resistors throughout) results in a high overall RSSI measurement accuracy, from antenna to bits. There sometimes is a case to be made for using more than one voltage reference circuit within the confines of a single IC. These cells are invariably quite small, and the isolation resulting from using separate cells is valuable. But these situations generally arise in less-critical systems. For example, in extensive tracts of current-mode logic, local cells are used for biasing.

In some cases, an even simpler solution is possible. This is the use of the raw supply voltage to scale both the RSSI function and the ADC (Figure 2.4). This approach is used in both the AD606 (a Log-Limiting IF Strip) and the AD608 (a Single-chip Superhet Receiver with Log-Limiting IF Strip). The RSSI output is scaled directly by the raw supply voltage, but this is also used by the ADC as its scaling reference. Thus, both bandgaps have been completely eliminated, with no loss of accuracy, as well as their supply current, die area, bonding pads, package pins and attendant ESD concerns, and guaranteed robustness. The only trade-off in this case is only that the components must be used in partnership. This slight loss of flexibility is never of great concern in high-volume system-oriented products.

2.4.2. Choice of Technology and Topology

Early in the design planning, we will select an appropriate technology for an IC product, based on issues of target cost, performance objectives, production capacity, time to market (and the possibility of cell re-utilization) and other issues of a strategic nature. In some cases, we will have little choice but to use a foundry process. We then start looking for robust circuit topologies – structures which have demonstrated low sensitivities to the absolute value of the individual passive components (minimizing the DAPs), and low sensitivities to mismatches, supply voltage and temperature (TAPs and STMs). The design principles are invariably the same: lean heavily on the use of ratios wherever possible, in the pursuit of TAPs; adopt sensitivity analyses and chose low-sensitivity cells in the case of DAPs; use careful layout techniques to address the STMs.

A couple of examples of techniques that address robustness will be presented. In the second of these, we will consider a rudimentary voltage-mode amplifier based on a pair of bipolar transistors with resistive loads. Open-loop amplifier cells of this sort are often deprecated, partly because of concerns about gain accuracy. Rather, the common tendency is to appeal to the use of op-amp techniques, in the belief that they automatically circumvent such problems, and conveniently transfer the attainment of high gain accuracy to the ratio of just two resistors. Occasionally, this may be effective, if the op-amp has sufficient open-loop gain at the frequency of operation. But this is often not the case in practice.

Indeed, one of the worst analog-circuit myths is the notion that the chief value of an op-amp is its "very high open-loop gain". Suppose we have an op-amp cell that has been proven to have a reliable DC gain of 10^6 and a *nominal* unity-gain frequency of 200 MHz, and we are planning to use this cell to realize an amplifier having the (seemingly low) numerical gain of ×12 at 10 MHz. We choose the feedback ratio accordingly. For an inverting configuration, the input resistor R_1 might be $1\,\mathrm{k}\Omega$ and the feedback resistor to the summing node R_2 would be chosen as $12\,\mathrm{k}\Omega$. With robustness in mind, we might decide to make R_1 as 3 units of $3\,\mathrm{k}\Omega$in parallel and R_2 as 4 units of $3\,\mathrm{k}\Omega$in series (Figure 2.5), use a generous width, and make sure the layout designer puts these resistors side by side, even interdigitates them and adds dummy resistors at each end to further ensure the ratio accuracy.

Then, in simulation (or perhaps in a bench experiment) we find that the actual gain is much lower; instead of ×12 it is found to be only ×9.6. Why? Because the open-loop A_{OL} gain *at frequency* is only 200 MHz/10 MHz, or merely ×20,

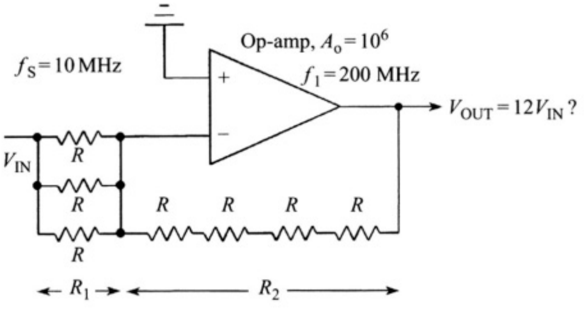

Figure 2.5. A fixed-gain amplifier designed to be robust: but is it?

assuming the usual case of dominant-pole compensation. At this juncture, one might decide to just make a correction to R_2, of slightly more than the wanted-to-actual gain ratio 12/9.6, to compensate for the lower A_{OL} at 10 MHz. Either through the use of vector arithmetic or simulation, we find that R_2 needs to be raised to 16.44 kΩ. This is no longer a low-integer ratio, but we choose to now use a total of *five* units for R_2, extending the length of each element by 9.6%, from 3 to 3.288 kΩ. A small change in the length (keeping the width constant) will not seriously jeopardize the ratio, because this dimension will invariably be relatively large. For example, using a sheet resistance of 1 kΩ/square and a width of 10 μm, the length increases from 30 to 32.9 μm, (the nearest 0.1-μm increment, resulting in an error of +0.06%).

We may think we are pursuing a sound "TAP" approach in using these "ratio-based" tactics, but this would overlook the important fact that the unity-gain frequency f_1 of the op-amp is itself a "DAP", being subject to variations in the on-chip resistor that determines the bias current and thus the g_m of the input stage, and variations in the on-chip capacitor; together these set the unity-gain frequency, f_1, which can easily vary by up to ±40% in production. Therefore, a *one-time adjustment* to the resistor ratio cannot guarantee accurate closed-loop gain at 10 MHz over all production units. In fact, the gain will vary from ×10.3 to ×13.2 over the lesser f_1 range of 150–250 MHz, a variation of only ±25% (Figure 2.6).

There are several ways in which this particular problem might be solved in practice. The preferred solution, whenever one has control over the complete ensemble, is to lower the op-amp's internal compensation capacitor and substantially raise the f_1 to a value better suited to the specific application of the amplifier cell, which no longer requires it to provide HF stability at all gains down to unity. Another solution, chosen to illustrate how robustness can often be achieved by the use of like effects, is shown in Figure 2.7. A second on-chip

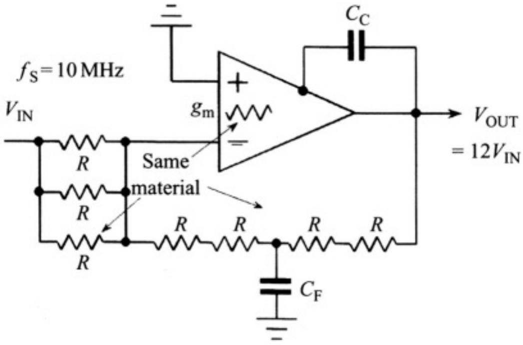

Figure 2.6. Improved amplifier using carefully-scaled bandwidth compensation.

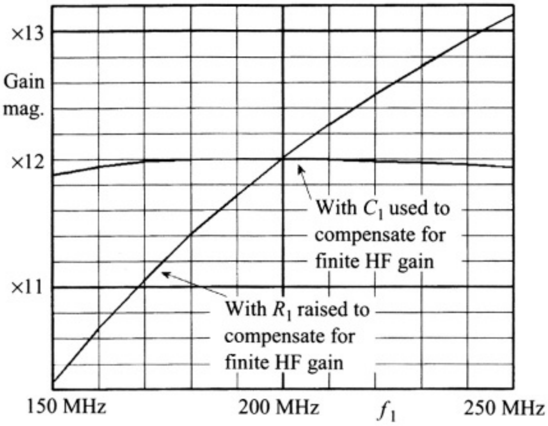

Figure 2.7. **Simulation results for original and improved amplifiers.**

capacitor C_F has been added at the junction of the two halves of R_2. If we make this component out of the same units as the internal HF compensation capacitor and also make the resistor that sets the bias, and thus g_m, of the input stage out of the same material and similar-sized units as R_1 and R_2, we can achieve a useful desensitization of the closed-loop gain at the presumed signal frequency of 10 MHz. Now we are *matching time-constants,* and on the path toward a true TAP situation.

It is useful to show how this improvement in robustness is obtained. We begin by modeling the op-amp's forward gain as that of an inverting-mode integrator $-1/sT_1$, and define the feedback time-constant as $T_F = R_2C_F/2$, and the magnitude of the closed-loop DC gain as $G_0 = R_2/R_1$. The transfer response of this circuit is

$$G(s) = \frac{-G_0\,(1 + sT_F)}{1 + (1 + G_0)\,sT_1 + (1 + G_0)\,s^2T_1T_F} \tag{2.2}$$

This is a two-pole response with a Q of T_F/T_1. Thus, we can rewrite (2.2) as

$$G(s) = \frac{-G_0\,(1 + sQT_1)}{1 + (1 + G_0)\,sT_1 + (1 + G_0)\,s^2QT_1^2} \tag{2.3}$$

It is easy to hold the ratio Q to within fairly narrow limits, since both T_1 and T_F are generated by CR combinations having exactly the same process sensitivities. For operation at a specific frequency, such as 10 MHz in this example, our only remaining concern is the *absolute value* of both time-constants, represented in (2.3) by the single integrator time-constant T_1. If we were concerned with the broadband response, we would choose to use a low Q; but since the main objective in this illustrative example is presumed to be the desensitization of $G(s)$ to the actual value of T_1 over a narrow frequency range, we may find it beneficial to use a somewhat higher Q.

Suppose we decide to make the magnitude of the gain $G(s)$ at the operating frequency equal to the target (DC) gain G_0. Solving (2.3) for Q we obtain

$$Q = \frac{1 + G_0}{1 - (1 + G_0)\,sT_1} \tag{2.4}$$

Thus, for $sT_1 = 10\,\text{MHz}/200\,\text{MHz} = 0.05$ and $G_0 = 12$, the optimal value of Q is 32.5. From (2.4) it also follows that this compensation scheme cannot be used above

$$sT_1 = \frac{1}{(1 + G_0)} \tag{2.5}$$

For the target gain of ×12, this technique can provide accurate compensation only up to 200MHz/(1 + 12) = 15.4MHz, at which frequency the Q would be dangerously high.

We might also determine the sensitivity to the value of T_1, and set that to zero. One could spend a few hours in this sort of analytical wonderland, but it would not be very helpful in providing practical insights. It is often the case that the actual operating conditions differ from those assumed at the start of a project, and all the effort poured into a specific analytical solution needs to be repeated. A more efficient way to explore the general behavior of such compensation techniques is invariably through *creative simulation.* The results that were shown in Figure 2.6 required about a minute of experimentation and optimization in real time (the maths shown above took considerably longer to go through). They demonstrate that, with the optimum choice of C_F and a small adjustment to R_2, good stability in the magnitude of the gain at 10 MHz (+0/–1%) is possible over a ±25% range of f_1, which represents the bulk of the yield distribution of a production op-amp.

In this brief exercise, we were able to convert troublesome DAP behavior into a benign TAP form; that is, we ensured an accurate gain at a significant fraction of the op-amp's unity-gain frequency, with a near-zero sensitivity of gain to that parameter at the chosen frequency. Even when a higher f_1 is employed, which, as noted, would be the preferable solution to minimizing this sensitivity, the addition of C_F would still be useful in further improving robustness in production, and at very little cost in die area, and at no cost in power consumption. By contrast, solutions based on further increasing the op-amp's f_1 *will* incur power penalties, within a given technology.

An excessive reliance on small-signal modeling with linear equations, and the use of small-signal simulation, is always a *very* risky business. Unfortunately, these methods are widely used in many theoretical treatments of circuits found in the academic literature to the neglect of the consequences of variations in circuit dynamics caused by perturbations in the working point, a result using signals of practical magnitude. Small-signal analyses and simulations totally hide numerous such effects.

It is common for device nonlinearities to introduce gain variations of a significant fraction of a decibel over the voltage (or current) swing corresponding to the full output of the circuit. This is the domain of nonlinear dynamics, which is invariably intractable using standard mathematical tools, while posing no problems to a simulator. Thus, one should spend relatively little time using simplistic frequency sweeps ("Bode plots") examining the gain magnitude and phase at some nominal bias point, and *far more time* in various kinds of dynamic sweeps. These include full transient simulations, pushing the circuit to confess its secret weaknesses, not only for comfortable operating conditions, but also at the extreme limits of the process, voltage and temperature (PVT) range, with comprehensive package models,[16] for worst-case source and load impedance, and the like. This issue is revisited in Section 2.5.7.

2.4.3. Remedies for Non-Robust Practices

One of the most intensively studied design topics is that of active filters, of both continuous- and discrete-time types, reflecting their importance in all fields of electronics. The better texts on the subject emphasize the need to choose topologies and/or component values *that formally* minimize the sensitivity of the dimensionless specifications, such as gain and the geometric disposition of poles and zeroes. Unfortunately, these same authors often show a poor appreciation of the need to convert a beautiful "minimum-sensitivity" design (in the strictly mathematical sense) into a practical, *manufacturable* entity.

For example, there is little point in concluding that the optimal (least-sensitive) solution is one in which, say, resistors of 5.3476, 1.0086, 1.7159 and $8.1030\,\mathrm{k}\Omega$ are needed, along with capacitors of similarly exotic values. Such component precision can rarely be met even in a board-level design. The chief appeal of text-book filter functions, such as the well-known Bessel, Butterworth and Chebyshev formulations, is simply that they are *mathematically tractable* and enjoy a certain sort of canonic rigor. But in these days of very fast computers and simulators, there is no compelling reason to stick to classical forms.

[16] It is essential to keep well in mind that circuits do not know what they're *supposed* to do, and design mastery entails making sure that transistors dance to *your* tune, not theirs. Thus, if you are using a common 25 GHz IC process to realize, say, an audio amplifier at the tail end of a receiver, the circuit will surely promote itself to a microwave oscillator, unless you pay attention to easily-forgotten parasitic effects having no essential relevance to your intended application.

The art of designing *manufacturable* filters begins with the sure expectation that some slight departure from the "ideal" (often over-constrained) response will be forced by the difficulty of actually realizing non-integer element ratios to high accuracy in a production context, and that it will be necessary to juggle the partitioning and topologies so as to force a solution using *simple integer ratios of Rs and Cs.* In modern filters, this paradigm is less often practiced in the design of continuous time filters than in switched capacitor filters. The most likely explanation is that the former were developed in the age of electrical theory, while the latter arose in an intensively pragmatic context, where it was known from the outset that unit replications would be essential to robustness.

The approach to monolithic filter design thus *starts with a trade-off,* namely, the need to set aside the text-books, and cut loose from the canonic rigor presented in the filter design literature. The ensuing design exercises may involve a considerable amount of "inspired empiricism" using the simulation of cells containing only element ratios that one knows can be reliably reproduced in high-volume production. Such an approach is straightforward for low-order filters, but can quickly become very difficult when advanced filter functions must be provided. However, in such cases, it is usually possible to create some adjunct routines to perform algorithmic optimization in a few minutes of unattended computer operation. Because filters are invariably required to be linear, the computational burden can be greatly simplified by the temporary use of idealized active elements in SPICE, or the use of a platform such as MathCad.

It should be realized that this is just a starting point. and it is important to note that an appeal to empiricism should not be confused with guessing, or even worse, lazy-mindedness. It simply recognizes that situations often arise in which systematic and analytic methods are either inadequate to the task at hand, or become too cumbersome to provide the needed rate of progress in a product development, or fail to generate insights that can be translated into practice. After empirical methods have pointed the way forward, it remains the responsibility of the designer to ensure a controlled and predictable outcome in the face of production tolerances. Empirical searches for manufacturable solutions are in no way a *substitute* for robust design based on fundamental considerations, but they are needed to explore the use of (and the invention of) more robust cell structures. Diligence will always be needed thereafter to preserve low sensitivities and reproducibility.

Some analog cells are inherently robust while others, that may quite appear similar, are not. Figure 2.8 shows two translinear multiplier cells.[17] The (a) form

[17] See B. Gilbert, "Current-mode circuits from a translinear viewpoint: a tutorial," in: C. Toumazou, F. J. Lidgey, and D. G. Haigh (eds), *Analogue IC Design: The Current-Mode Approach,* Chapter 2 IEE Circuits and Systems Series, vol. 2, Peter Perigrinus, 1990.

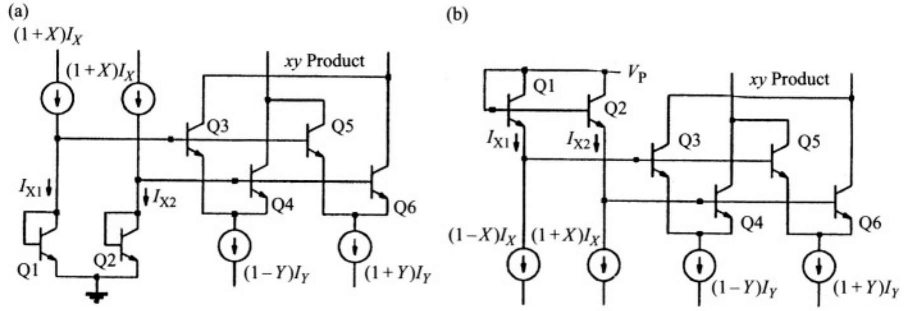

Figure 2.8. Two fundamental multiplier cells; (a) beta-immune, (b) beta-prone.

is called “beta-immune”, because its scaling is very little affected by BJT current-gain, β, and can remain accurate even when I_Y is almost as high as βI_X. The (b) form is called “beta-prone”, because its scaling is sensitive to beta, even for much less demanding bias conditions, for example, when $I_Y = I_X$.

The explanation is straightforward: in (a) all the base currents in Q3–Q6 are in phase with the corresponding currents in Q1 and Q2, and the ratios of I_{E3}/I_{E4} and I_{E5}/I_{E6} remain strictly equal to I_{E1}/I_{E2}. Assuming the betas are essentially equal and independent of current, the input-linearizing transistors are not affected by the reduction in the *absolute* bias levels due the current robbed by the bases of Q3–Q6, because these are in exactly the same ratio as I_{X1}/I_{X2}. On the other hand, in the (b) cell, the base currents are out of phase with the inputs, and the ratio I_{E1}/I_{E2} is therefore not equal to the input-current ratio. The overall consequence is that the scaling of the (a) cell includes the factor $1/(1 + 1/\beta)$, while for the (b) cell this factor is approximately $1/(1 + 3/\beta)$. Here we have a good example of a *trade-off in topology.* In practice, the (b) form is easier to drive (from voltage-to-current converters using the same device polarities for both the *X* and *Y* signals) than the (a) form, and the literature shows that the (b) cell has almost universally been chosen in monolithic analog multipliers because of this topological advantage, at the expense of static accuracy, temperature stability, intermodulation and slightly higher noise (due to the base currents of the core transistors). However, the beta-dependent scaling can be easily compensated in the synergistic design of the associated voltage reference.

2.4.4. Turning the Tables on a Non-Robust Circuit: A Case Study

This lesson underscores the general point. Good topologies and biasing practice are fundamental requirements in the pursuit of sensitivity minimization, in

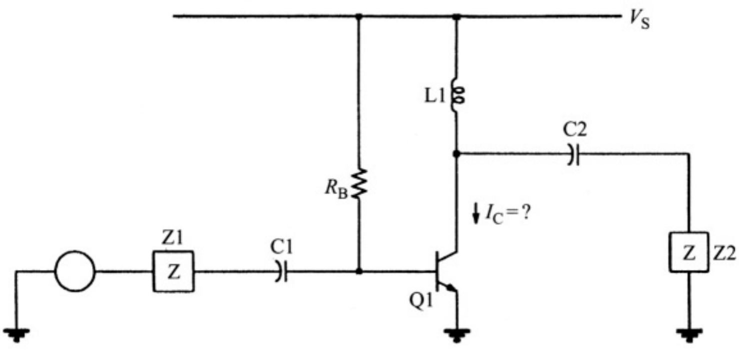

Figure 2.9. An LNA having inadequate attention to biasing.

the face of every sort of environmental factor, notably "P" (lot-to-lot production spreads leading to absolute parameter uncertainties), "V" (supply voltage) and "T" (temperature). In numerous cases, we would have to add "M" (matching) as one of these factors, though not in the following case study. Figure 2.9 shows the circuit, a low-noise RF amplifier (LNA). The topology used here is open to criticism, although the form was once widely used. The behavior of this *one transistor* circuit can be surprisingly complex, and abounds with trade-offs and compromises. It is often nonchalantly presented in articles with almost total emphasis on its high-frequency aspects and hardly any on the crucial matter of choosing and regulating the bias point. It amply illustrates the peculiarities of analog design.

We will focus here on the biasing methods. The general method shown, and regrettably still all-too-often employed in discrete-transistor RF design, uses a high-value resistor R_B taken directly to the supply voltage, V_S, in order to establish the collector current I_C. This immediately introduces serious and quite unacceptable sensitivities, of at least four kinds.

First, we need to understand that the precise value of I_C, and its temperature shaping, affects all aspects of BJT performance, and thus that of the LNA. The g_m is essentially proportional to this current,[18] and the noise, power gain and input impedance are all dependent on it. This would be true even at moderate frequencies. At high frequencies these parameters are far more seriously impacted, since I_C also affects the f_T of the transistor; for this class of operation (and as a general rule) this will be much lower than the peak f_T, which occurs

[18] Neglecting for the moment the effect of impedances in the emitter branch. For example, inductance may be incidentally introduced by the bond-wires and package, or deliberately used to desensitize the g_m to I_C.

only over a very limited range, and at current-densities above those usually permissible.

This crude rationale was based on assumptions of this sort: (1) I_C must be 2 mA; (2) the nominal DC beta is 100; (3) the nominal V_S is 3 V. A base current of $20\,\mu A$ is, therefore, needed, and using reference data, is was found that the V_{BE} for the particular transistor type is 800 mV at this current. The nominal voltage across R_B is thus 2.2 V, and this resistor must be $2.2\,V/20\,\mu A = 110\,k\Omega$. Choosing an IC resistor layer which has a sheet resistance of $440\,\Omega/\text{square}$, it is found that 250 squares are needed. Not wishing this ("trivial") biasing component to be physically too large, one might decide to make it $2\,\mu m$ wide by $500\,\mu m$ long. Now, if we examine the numerous ways in which sensitivities have been carelessly introduced into this cell, we find:

1 I_C varies with the supply voltage, V_S, in a more-than-proportional way. Noting that $I_B = (V_S - V_{BE})/R_B$, the sensitivity is increased by the factor $V_S/(V_S - V_{BE})$, or about 1.36. Thus, over the range $2.7\,V \le V_S \le 3.3\,V$ the g_m will alter by about ±14%.

2 I_C is essentially proportional to the DC beta. Over an assumed worst-case range of $35 \le \beta_{DC} \le 200$, the collector current (and thus the g_m) would vary from one third to twice its nominal value.

3 I_C is extremely sensitive to the delineation of resistor width, chosen as $2\,\mu m$. If we suppose that the worst-case variance on this parameter totals $\pm 0.25\,\mu m$, the g_m will vary by ±12.5%. (We can fairly safely ignore the length variation in this case.)

4 Using this biasing scheme, the g_m will vary with temperature for several reasons. The DC beta will vary by typically +1%/°C; over the range $-55°C < T < +125°C$, this is a large effect. Also, the V_{BE} varies by roughly –1.5 mV/°C, causing I_C to increase by another 0.07%/°C. However, a resistor TC of 1,000ppm/°C would fortuitously lower the last two effects. Some polysilicon resistors, however, have a negative TCR, which will aggravate the sensitivity.

Now let's redesign this rudimentary, but important, basic cell with robustness uppermost in mind. We must begin by squarely facing these facts:

1 In BJT practice, the control of g_m is invariably of paramount importance. In this LNA it is a major factor in the determination of gain, noise figure and the accuracy of the input/output matching. Without inductive degeneration in the emitter, the sensitivity to g_m is maximal; even when such is added, some sensitivity remains.

2 The basic g_{m} is qI_{C}/kT – it is directly proportional to collector current. This is a very reliable relationship, the fundamental basis of translinear design, and remains true even when the signal frequency f_{S} is a substantial fraction of the f_{T}. It will be diluted somewhat by the presence of significant base resistance $r_{\mathrm{bb'}}$, and, in all modern transistors using polysilicon regions for emitter contacting (which includes SiGe structures), by the emitter resistance $r_{\mathrm{ee'}}$.

3 It follows that I_{C} must be proportional to absolute temperature (PTAT) if the g_{m} is to be stable; furthermore, this condition must be maintained in the presence of *unknown* values for $r_{\mathrm{bb'}}$ and $r_{\mathrm{ee'}}$.

4 Therefore, if the design is to be robust, I_{C} must have a low sensitivity to V_{BE} and β_{DC}; it must have a low sensitivity to V_{S}; and it must be desensitized with respect to the delineation of on-chip resistances. This last objective stems from the need to achieve accurate impedance matching at both the input and output ports of an LNA, but the need for resistor control is in conflict with production variances in sheet resistance as well as lithography. It must be noted that numerous extant designs give little attention to matters of this sort and the design process may use S-parameters throughout with no regard for the fact that these are but snapshots of the full reality, *relevant only to one particular bias point.*

5 In designing the associated biasing circuit, we will need to remember that I_{C} will be a function of V_{S} also through the effects of Early voltage, V_{AF}, since (in the non-robust design being considered here) its collector is taken directly to V_{S}, while its base–emitter port is close to ground. Furthermore, variations in f_{T}, C_{JC} and (except in silicon-on-insulator processes) C_{JS} will impact the HF performance. Therefore, any improvements must seek ways to minimize all these sensitivities.

Whatever design choices we make, we should instinctively strive to find *the simplest possible solution.* On the other hand, in a monolithic context, this does not mandate the sparse use of transistors. Since this chapter is not concerned with LNA design in general, we cannot afford to pursue this example as fully as it deserves, but we can address the above challenges to robustness with the following observations.

Item (1) touches on a broader issue of LNA design, namely, the use of reactive (noise-free) emitter degeneration to lower the sensitivity of the effective g_{m} (now defined by the vector sum of $r_{\mathrm{e}} = 1/g_{\mathrm{m}} = kT/qI_{\mathrm{C}}$, and the inductive reactance $Z_{\mathrm{E}} = \omega L_{\mathrm{E}}$). This also serves in the interest of robustness, because inductors can readily be fabricated on-chip, and have narrow production tolerances[19] being largely dependent on the number of turns. For present purposes, the Q of the inductor does not need to be high, and it may be made in spiral form using the aluminum interconnect. It will often have a few ohms of resistance; using a typical metal thickness this amounts to roughly one ohm per nanohenry, which means that the Q will be constant at 2π, or about 6.3.

This resistance will vary, due to variations in the thickness (hence, sheet resistance) of the metal, and the width of the spiral trace, which is subject to photolithographic and etching variances. However, its resistive component will form only a small part of the overall emitter impedance, and is of relatively little consequence to the determination of gain and linearity. We can expect that there will be further impedances in completing the emitter branch and connecting to the system (board-level) ground, a path that includes the bond-wire(s) and the rest of the IC package. The inductive components will be predictable, but, keeping robustness in mind, we will want to ensure that the method used for *biasing* is not sensitive to the addition of *unknown resistances* in the emitter branch.

Item (3) demands that the I_C be PTAT. Numerous cells are available to generate a PTAT voltage, based on ΔV_{BE} techniques. Through the use of an appropriate topology, this voltage can have a low sensitivity to V_S; we can even embed it in full compensation for the effects of V_{AF} on I_C in the LNA transistor. This voltage can be converted to a current through the use of a resistor. However, when this resistor is on-chip, we will not fully satisfy the last criterion in Item (4), but we can greatly reduce the sensitivity to photolithographic and etching variations by using a physically large resistor, leaving the unavoidable uncertainty in the sheet resistance of the resistor layer and its temperature coefficient.

In many theoretical treatments of circuit design, properties such as g_m are presumed to be inherent to the device, although dependent on the bias current, which is treated as a "merely practical" consideration. However, this viewpoint is ill-advised. The generation of reliable currents in an IC using bipolar transistors, and in some RF CMOS circuits, is based on the use of resistors, and the currents will therefore be poorly controlled when these are on-chip. One may be able to still achieve robust operation when this is the case, but generally speaking many of the properties of an analog IC, such as the terminal impedances of a broadband amplifier, are directly traceable to a real resistor somewhere on the chip, and the voltage that is imposed across this resistor.

[19] There is another very important reason why we will choose to use inductance in the emitter, which is in connection with intermodulation. The constant inductive reactance can be made much larger than the nonlinear r_e, thus greatly improving the linearity of the overall transconductance.

Not only are the port impedances a reflection of a physical resistor, but other parameters such as gain and bandwidth may be.

For example, if a wideband amplifier is constructed using a BJT differential pair as a transconductance stage, with resistive collector loads, the gain $g_m R_L$ can be stated in the form R_L/R_0, where R_0 is the transformed value of a bias resistor embedded in the chip. Thus, just as for a feedback amplifier, the gain magnitude is a simple, dimensionless ratio, which can be quite accurate when the relevant precautions are observed. For an op-amp, whose open-loop gain at a practical signal frequency f_S can be stated solely in terms of its unity-gain frequency, f_1 (its magnitude is f_1/f_S),the situation is more complex, since f_1 is determined by a CR product, and cannot be accurate without trimming. Only at very low frequencies does an op-amp's gain become a simple ratio, and then a rather uncertain one.

Holistic optimization of the LNA. It is apparent that we have been pursuing a "whole↔part↔whole" approach to this LNA, which is essential in the relentless pursuit of robustness. We started with the *whole* circuit (usually not so simple!) and then moved in closer to think about just *apart*: the biasing details. In doing so, our attention was eventually directed back to the *whole* again: the search for a new topology. It was not essential to respond to this undercurrent of concern about biasing. We could have stuck doggedly with the whole – the original circuit. But in considering how to improve the biasing part, we came to realize that this was actually a crucial and multi-faceted question, and that the flaws in the original topology were deep, necessitating a search for ways to improve the whole design, not just "choose the bias point". (Although we cannot pursue the topic here, there is a formal optimization of the bias related to the minimization of noise figure, but that does not overshadow the above considerations.)

While we fully expect to use simulation to fine-tune this design, particularly with regard to its two-port characteristics at, say, 1.9 GHz, and with the full package model included,[20] it is difficult to see how we could hand over the challenge of producing a robust design to some kind of optimization procedure. Such a program can be no better than its writer in foreseeing all the myriad ways in which a handful of components can be connected to make an improved cell. This is clearly not a matter of simply instigating an automatic search of all possible solutions and then evaluating them all to find the "best" one, in basically the same way that Deep Blue wins at chess. In the first place, we would have to decide on some very simple constraints, such as the maximum

[20] Which, in addition to the simple series inductance of bond-wires is also rife with other parasitics, including mutual inductance between these wires, the effect of which is difficult to quantify without simulation studies.

number of components that allowably could be used, and their mix (so many transistors, so many resistors, etc.). But more importantly, the critical value of certain spontaneous, unforeseen and creative topological alterations will necessarily be overlooked in a finite procedure. Even a skilled designer-programmer could not anticipate every combination and every consequence needed to drive a branching heuristic. An appeal to random topological variations would generally lead to nonsense. Equally problematical, one needs to formulate elaborate and all-embracing evaluation functions, "goodness" criteria that tell us when we are getting closer to a "better" solution, however, that may be defined.

Given a very limited set of performance criteria and only a few permissible topologies, some useful optimization may be possible in this way. However, the benefits to be gained from such a program would need to be weighed against the time taken to write it, and the number of times it would be used.[21] Such projects invariably fail, because they do not provide enduring practical value. A more serious criticism of the "Optimizer" approach to product development is that it may be seriously misunderstood by young designers, who are inclined to use "clever programs" of this sort rather than confront what seems to be the formidable challenge of learning the individual details attendant to each class of circuit. The allure of quickly having results in hand, no matter what is inside the program, may be hard to resist.

To return to our LNA, we can already see changes are going to be needed in the biasing, and perhaps in the topology, too. There is also the challenge of choosing a close-to-optimal size for the transistor, where we will be confronted with more trade-offs. Since the effective g_{m} is influenced by $r_{\mathrm{bb'}}$, we will choose a device geometry that minimizes this parameter as far as practicable before the f_{T} of the transistor begins to suffer appreciably due to the reduction in current-density and the increase in junction capacitances. We also need to minimize $r_{\mathrm{bb'}}$ in order to achieve an acceptable noise figure. However, large devices will have a high C_{JC}, which, in the prototype topology, will have a low capacitive reactance at high frequencies. This is a very common trade-off in RF design. Knowing that the C_{JC} of a large, low-noise transistor may be high, we might

[21] The writer speaks from painful experience. In 1960, using an Elliott 803 vacuum-tube computer, he wrote a program for *The Automatic Design of Circuits.* It really did what it claimed, for a small set of circuits. It selected the best devices for a given application out a library of 36 germanium transistors, and given simple boundary objectives, such as gain, input noise, bandwidth and the like, it calculated all component values, later selecting the nearest available standard values and recalculating the bias point and the subsequent effect on the terminal performance. It then carried out a specified number (the default was 1,000) of Monte-Carlo analyses and predicted board (not chip) yield for typical production variances and their correlation factors. This labour of love was used once or twice, in a serious capacity, and some examples published in the professional literature. Then, it fell into oblivion.

start thinking about the use of a cascode transistor to minimize the impact of this capacitance on the feedback impedance Z_F, which we will use to more reliably control the LNA parameters. A cascode is also consistent with the need to reduce the effect of the supply voltage.

But this entails another trade-off, namely a reduction in the available voltage swing at the collector and/or a tightening of the constraints on supply voltage. It must also be remembered that the emitter impedance of the cascode transistor at high frequencies is not the simple resistive $r_e = kT/qI_C$; rather, it is markedly inductive. A yet further trade-off then arises: a small transistor is needed here, to reduce its capacitances C_{JC} and C_{JS}. Large values would not only complicate the output matching but also introduce even-order distortion due to their varactor behavior; and there is a further subtle source of noise, often overlooked, arising from the resistance associated with C_{JS} and its Johnson noise. A large C_{JE} in this device will cause further parametric distortion.

So, we decide to use a small transistor. But this will have a high R_C, and at the currents often used in LNAs, this will reduce its collector junction voltage, perhaps almost to the point of saturation at the most negative signal swings at its collector. This region of operation will lead to further distortion and intermodulation. (Here, we are once again in nonlinear territory.) Furthermore, its high $r_{bb'}$ is transformed in its emitter branch – the collector load of the large transistor – into an inductive component, leading to further effects in the overall behavior of the LNA. Thus, the scaling of the cascode involves several other trade-offs.

The next step will be a pencil-and-paper session of sketching out a few other topologies, to consider different ways the trade-offs may be resolved. This sort of exercise comes easily to the experienced designer, who is unlikely to be in awe of the well-established approaches, and who realizes that the aggressive exploration of numerous alternatives is an essential part of the design process, often leading to valuable new insights, even breakthroughs which later become classics in their own right. Designing involves traveling down many dead-ends. It is as much about discovering or devising new cell forms as it is about simply “choosing the bias point”, calculating a few component values, and performing perfunctory simulations to offer at the Design Review as a smoke screen to distract from the absence of real invention.

After a brain-storming session of this sort, we may find that a circuit like Figure 2.10 offers a pretty good fit to the circumstances. The supply-insensitivity is achieved through the use of an adjunct bias cell, which for the time being we can describe as a band-gap reference, generating $E_G = 1.22\,\text{V}$ independent of the supply. This provides the bias for the base of the g_m device, Q1, now delivered through a moderate-sized, and therefore more controllable, resistor, needed only to block the RF input from the bias cell. The emitter current is then defined by the resistor R_E, which we can choose to put either

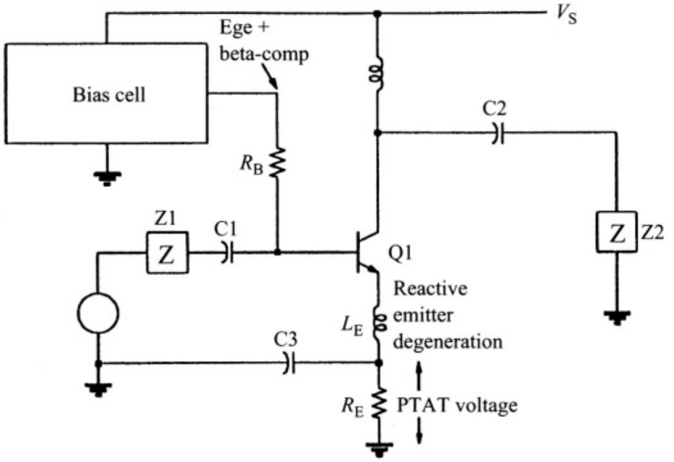

Figure 2.10. Revised LNA, having precise bias control and greater robustness.

off-chip or on-chip. An extra pin is not required, since this path to ground must already be separated from generic power-ground pins, so that choice will mainly depend on the required accuracy of gain and impedance matching. Since the bulk of the emitter impedance is now determined by the inductance L_E, using a sufficiently high value of I_C so that $r_e \ll Z_E$, we have largely desensitized the gain and matching to variations in the bias current. A fairly high current is needed anyway to achieve a low input-referred voltage-noise spectral density due to shot noise mechanisms; the input-referred voltage is proportional to $1/\sqrt{I_C}$ and evaluates to $0.27\,\mathrm{nV}/\sqrt{\mathrm{Hz}}$ at 3 mA.

The sum of a suitably-scaled PTAT voltage and a V_{BE} can be made equal to E_G, the so-called band-gap voltage. Here, the reverse principle is being applied: since we are applying E_G to the base, the voltage across R_E, that is $E_G - V_{BE}$, will be PTAT, and thus so will I_E when R_E is a zero-TC resistor, as would be basically the case when this resistor is placed off-chip.[22] Another benefit that accrues from the use of resistive biasing in the emitter is that the sensitivity to the collector–emitter voltage is also lowered, over that of the first LNA. Taken alone, this consideration eliminates the need for a cascode transistor (whose base would be taken to a regulated voltage of about one V_{BE} above E_G or roughly 2 V above ground), to the extent that it serves to decouple supply variations from the collector of Q1. We can afford to omit the cascode

[22] A full discussion of biasing techniques is out of place here. However, we may mention that special methods can be used to generate PTAT currents using resistors of non-zero temperature coefficient.

on these grounds, but may still decide to include it when the high-frequency response is considered. When we do, we will have to revisit all those trade-offs.

The bias cell used to generate E_G could just be some previously-designed band-gap reference. But there is no need to set the bias voltage V_B to E_G, since this voltage does not need to be stable with temperature. It is only necessary to make it the sum of a V_{BE} (tracking the V_{BE} of isothermal Q1) and set up a PTAT voltage V_{PT} across R_E. We could choose to make V_{PT} as high as possible, in order to minimize the effect of errors due to mismatches in the V_{BE}s or arising across base resistors. Again, we are faced with a trade-off, since the higher bias voltage will erode the available voltage swing at the output, lowering the 1 dB gain-compression point. In such cells, it is an easy matter to include a 'beta-fix' in the bias voltage, to compensate for the finite DC beta of Q1, ensuring that at least its bias is accurate, although there remains an unavoidable sensitivity to the AC beta, which is approximately f_T/f_S; for an operating f_T of 12 GHz and a signal frequency f_S of 2 GHz, this is only 6. This is only one of several key parameters that are in the nature of "DAPs", and which unavoidably determine into the overall performance; in fact, there are very few "TAPs" in an LNA.

Integrated with the LNA, the optimized biasing scheme might look like Figure 2.11, in an all-NPN design. LNA designs of this sort can nevertheless provide acceptably accurate gain (± 1 dB) and matching (return loss > 15 dB) at high frequencies, with low sensitivities to supply voltage, temperature, current-gain and Early voltage. That is, they can be rendered *more* robust by careful attention to biasing issues, and the use of *synergism* in the biasing cell. Clearly, there is much more to robust LNA design than can be presented here and these comments are offered only to illustrate the sort of considerations that must be applied. It is noteworthy that *"trade-off"* occurs over ten times in this section

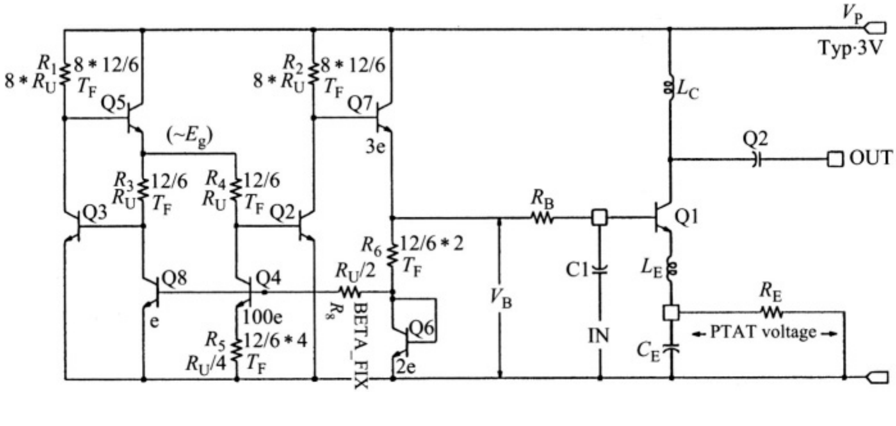

Figure 2.11. A more carefully crafted bias generator integrated into an LNA.

alone, concerned with a *one-transistor circuit* and in almost all cases, the context is not that of a pair-wise selection. This hints at the complexity of the trade-offs that must surely be expected of more typical analog circuits.

A further example of biasing synergy. Techniques of this sort – in which robust performance is ensured through the *progressive and systematic* elimination of sensitivities – are of central importance in analog design. In the next example, we will use a different approach to desensitize the gain of an open-loop amplifier in the presence of large variations in *junction resistances.*

Figure 2.12 shows a rudimentary gain cell based on a differential bipolar pair. The "simple-theory" unloaded small-signal voltage gain is

$$G_O = \frac{I_E R_C}{2V_T} \tag{2.6}$$

The first point of note is that this is another example where one would not choose to use a temperature-stable bias current. Rather, just as for the LNA, the collector currents must be basically PTAT to ensure temperature-stable gain. This is the general rule and is sometimes thought to be the only correct choice. PTAT biases are readily generated using a ΔV_{BE} cell, which generates some multiple of $V_T = kT/q$, let's say σV_T. In some way, this voltage is converted to a current by a resistor R_G. Thus we can rewrite (2.6) as

$$G_O = \frac{\sigma V_T}{R_G}\frac{R_C}{2V_T} = \frac{\sigma R_C}{2R_G} \tag{2.7}$$

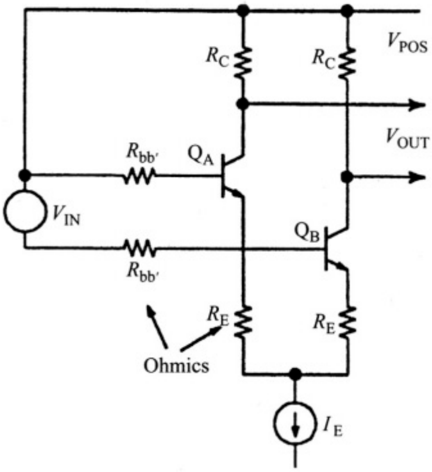

Figure 2.12. A rudimentary gain cell.

Now we appear to have a pure ratio. For small bias currents, the gain will be quite close to the theoretical value, up to fairly high frequencies, except for a small error due to the finite current-gain, β_{DC}, which can easily be corrected. But at higher values of bias (lower values of R_C and R_G), we will find the gain to be lower than expected. That is, we apparently have a DAP situation, *even though we thought we were invoking strict ratios.* Why?

It doesn't take long to realize that the finite junction resistances are responsible. Both the base resistance $r_{bb'}$ and the emitter resistance $r_{ee'}$ are involved. It is convenient to refer all such effects to the emitter, modeled in this figure by the resistors $R_E = r_{ee'} + r_{bb'}/\beta_{DC}$. Figure 2.13 shows the resulting gain error versus $V_E = I_E R_E$, in units of V_T. For example, when $R_E = 5\,\Omega$ (due, say, to $r_{bb'} = 200\,\Omega$, $\beta_{DC} = 100$ and $r_{ee'} = 3\,\Omega$) the gain error is — 8.8%, or

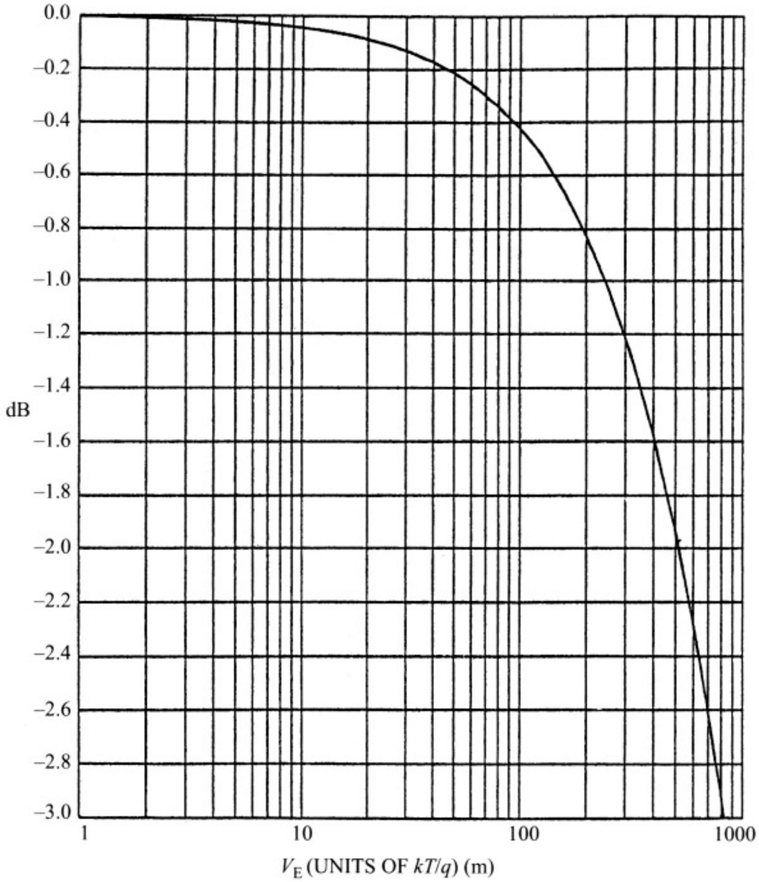

Figure 2.13. Gain error in the basic cell due to ohmic resistance.

–0.8 dB, at $I_{\mathrm{E}} = 1$ mA. Clearly, this error will vary from one production lot to the next, and appears to be a basic flaw, involving an unavoidable dependence on absolute parameters: the transistor junction resistances. The obvious, "brute-force" solution is to increase the size of the transistors so as to lower these resistances, but this route represents an unacceptable trade-off when the maintenance of a high bandwidth is another goal of the design. Similarly, the use of a lower I_{E} and higher R_{C} will likewise lead to a loss of bandwidth.

In a family of ICs now in high-volume production, it was essential to push the bandwidth out to about 4 GHz, and neither of the above solutions could be used. However, there is a very simple way to virtually eliminate this error, entailing only the correct design of the bias cell, *with no added components and no trade-offs* in either gain accuracy or bandwidth. This being the case, it might as well be employed as a matter of routine to improve the robustness of the design. In fact, this proprietary technique[23] is valuable even where much lower bandwidths are required, as in IF amplifiers. We will not discuss here the techniques by which the linearity can also be improved to well beyond that of the simple BJT differential pair, as these touch only indirectly on the robustness theme.[24]

Such corrections are possible because we can view this cell as an *analog multiplier,* whose gain is essentially proportional to I_{E}. Through the careful crafting of this current, a variety of subtle effects can be introduced, including the desensitization to both resistance and to beta. Putting aside the second of these errors for the moment, we can write the *actual* gain as

$$G_{\mathrm{A}} = \frac{I_{\mathrm{E}} R_{\mathrm{C}}}{2V_{\mathrm{T}}\,(1 + I_{\mathrm{E}} R_{\mathrm{E}}/2V_{\mathrm{T}})} \tag{2.8}$$

which is significantly in error when $I_{\mathrm{E}} R_{\mathrm{E}}$ is comparable to $2V_{\mathrm{T}}$. The junction resistance R_{E} depends on the size of the transistors used in the gain cell. Let R_{Ω} be the effective emitter-referred junction resistance of a "unit" device, and assume that the gain cell transistors use N unit emitter–base regions. Then, $R_{\mathrm{E}} = R_{\Omega}/N$. Using (2.8), we can readily calculate the actual value of I_{E} required to correct for the gain error:

$$I_{\mathrm{EA}} = \frac{I_{\mathrm{EO}}}{1 - I_{\mathrm{EO}} R_{\Omega}/2N V_{\mathrm{T}}} \tag{2.9}$$

This at first appears to be an awkward function to implement, but in fact, it readily can be achieved when the associated bias cell is considered as an *integral*

[23] B. Gilbert, US Patent 4,929,909, *Differential Amplifier with Gain-Compensation,* issued May 29,1990.

[24] However, the interested reader is referred to "The multi-tanh principle: a tutorial overview", *IEEE Journal of Solid-State Circuits,* vol. 33, no. 1, pp. 2–17.

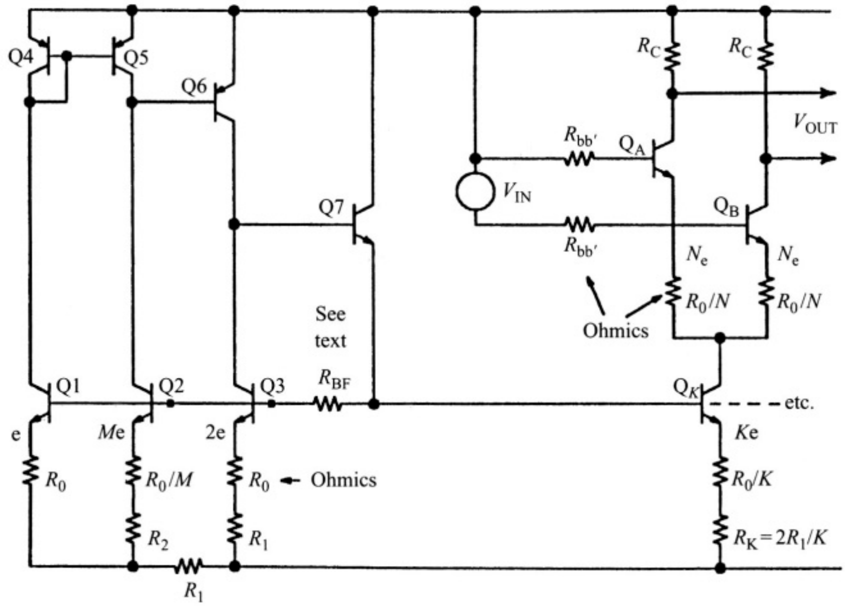

Figure 2.14. **Multi-stage amplifier with a synergistic bias cell.**

part of the design. Once again, we are seeking a *holistic solution* in the interest of minimizing sensitivity. Figure 2.14 shows a representative scheme.

For the moment, ignore the resistor R_{BF}. This figure also shows the junction resistances associated with Q1 and Q2 in the ΔV_{BE} cell. The *baseline* value for the currents $I_1 = I_2 = I_{CO}$ is just $(V_T/R_2)\log \mathrm{M}$, but the *actual value* is

$$I_{CA} = \frac{I_{CO}}{1 - (R_\Omega/R_2)(1 - 1/M)} \tag{2.10}$$

Note the similarity in the form of (2.9) and (2.10); it beckons us to equate the denominators, and thus eliminate the dependence on R_Ω. The required condition is

$$\frac{I_{EO}R\Omega}{2NV_T} = \frac{(R_\Omega/R_2)}{1 - 1/M}$$

Noting that $R_2 = (V_T \log \mathrm{M})/I_{CO}$ and that, in general, I_{EO} is K times I_{CO}, we arrive at the condition

$$\frac{K}{2N} = \frac{(1 - 1/M)}{\log M} \tag{2.11}$$

This condition ensures that systemic variations in R_Ω will not affect the gain. But we have yet to find the value of R_2 required to set this gain to the required value. Assuming that (2.11) is satisfied, we can use the baseline ($R_\Omega = 0$) equations to do this. The result is

$$R_2 = R_C \frac{K \log M}{2G_O} \tag{2.12}$$

In a robust, manufacturable design, *N, K* and M should all be integer. It is also desirable to find an integer relationship between R_2 and R_C, allowing the use of unit resistor sections. Such convenient solutions may not always be possible, but a little manual iteration will often reveal a solution which is "almost-integer", needing only small adjustments to the length of resistors and thus maintaining a low sensitivity to absolute dimensions.

For example, beginning with a nominal gain objective of ×4 (12.04 dB) and choosing $R_C = 1\,\mathrm{k}\Omega$, the required I_{EO} is $206.8\,\mu\mathrm{AP}$ and a target value for I_{CO} of $\sim 100\,\mu\mathrm{AP}$ puts the required integer value of *K* at 2. Choosing $N = 4$ and solving (2.12) shows that a value of $M = 50.5$ is close to ideal. Then, in solving for R_2 using (2.12), one finds that it would need to be $980\,\Omega$, which is not quite integer to R_C. However, using the adjusted fully-integer solution $M = 55$, $R_2 = 1\,\mathrm{k}\Omega$, the gain is only 0.02 dB high for $R_\Omega = 0$.

Figure 2.15 shows that the gain error remains negligible for values of R_Ω as high as $100\,\Omega$, when the maximum resistance in the emitters of Q_A and Q_B is $25\,\Omega$, that is, 10% of the r_e. The lower panel shows the corresponding increase in I_E needed to effect this compensation. In the ongoing pursuit of robustness, we would complete the compensation of gain errors by turning our attention to the effects of the finite DC beta, β_{DC}, in both the amplifier and bias cells. The ΔV_{BE} cell generates accurate currents in its *emitter* branches, so while the current in R_K accurately replicates that in R_1, the collector current of Q_K, and thus the gain, is low by the factor $\alpha = 1 - 1/\beta_{DC}$. Further, the g_m of the Q_A, Q_B pair is determined by their *collector* currents, which are low by a similar factor. (This is not "counting twice".) By including the resistor R_{BF}, the bias voltage is raised by an increment that increases as beta falls.

Note as a matter of detail (that's analog design) that the beta of Q_K will increase with the supply voltage, while that of Q_A and Q_B, operating at a V_{CB} roughly equal to zero, is slightly lower and *not* supply-dependent. By placing R_{BF} in the position shown, the current in it, and thus the compensation voltage, reflects the beta of Q1 and Q2, whose V_{CB} increases with the supply in the same way as that of Q_K, while that of Q3, whose V_{CB} is fixed, tracks that of Q_A and Q_B. A simple calculation suggests that it should, in this case, be roughly equal to R_1, but a slightly higher value (here $1.33\,\mathrm{k}\Omega$) provides more accurate compensation at very low betas.

The gain error (Figure 2.16) is under 0.05 dB over an extreme range of the SPICE parameter BF (roughly β_{DC} for moderate injection levels and low V_{CB}); the sensitivity to supply voltage is under 0.005 dB/V for VAF = 100 V. In a

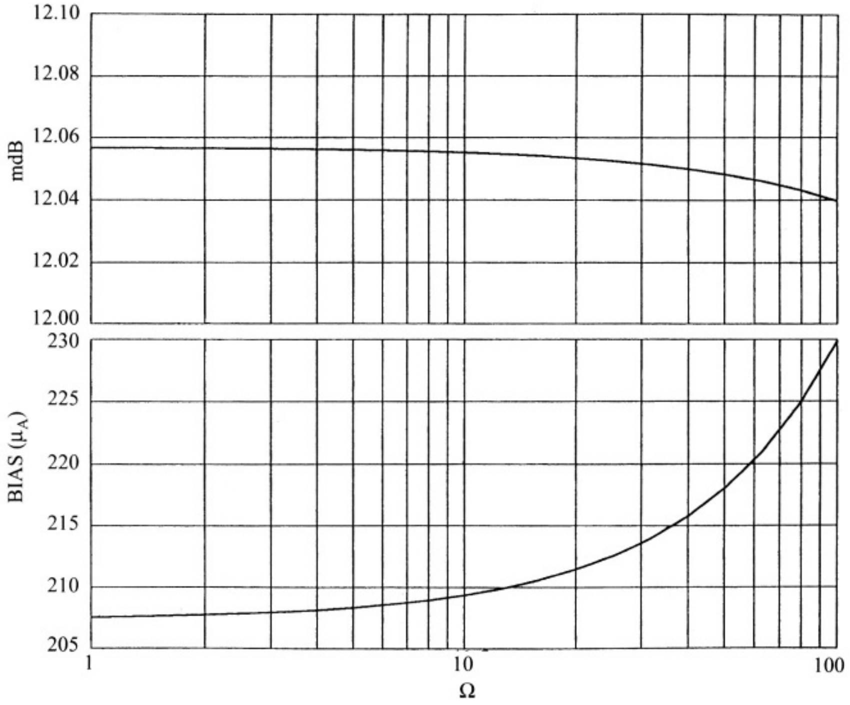

Figure 2.15. Cell gain error using synergistic bias technique.

multi-stage direct-coupled amplifier without the buffering advantage of emitter-followers between each cell, further gain errors arise due to the loading of the subsequent cell. This has a similar form, and can be closely compensated using a modified value for R_{BF}. The cell gain variation over the temperature range –55°C to 125°C is under 0.01 dB for this synergistic duo, further evidence that all the significant device variations affecting the mid-band gain have been addressed. The gain roll-off at high frequencies, while fundamentally of the nature of a DAP related to device inertia, can also be addressed in a synergistic and self-compensating fashion.

Biasing techniques of this sort can be applied to a wide variety of other errors in order to enhance manufacturability. With thoughtful use of optimal biasing methods, and sensible use of integer ratios of unit devices, very significant improvements in robustness can be assured, with little topological complication or the expenditure of more power. While the present examples are limited to bipolar studies, similar compensation methods based on assumptions of bias tracking can be applied to CMOS circuits. Indeed, even greater care is needed in this medium, where process variations are frustratingly high.

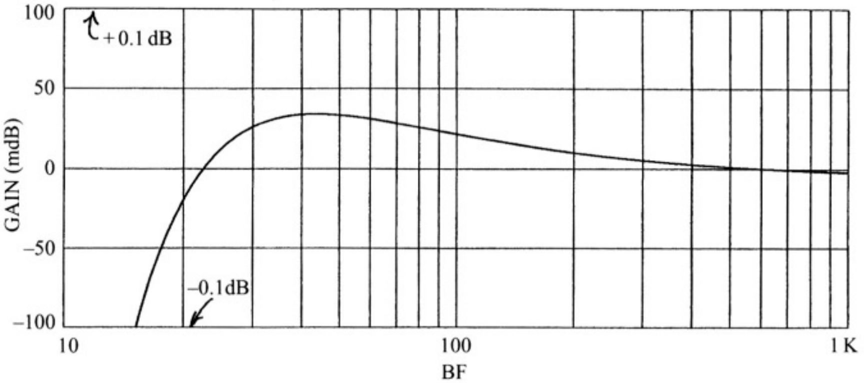

Figure 2.16. Cell gain error using beta compensation.

2.4.5. Robustness in Voltage References

There seems to a good deal of misunderstanding about the use of voltage references. Nowadays, the term “band-gap reference” is used very loosely. It is often applied to any cell in which a difference of junction voltages, that is, a ΔV_{BE} is used for general bias purposes. In this capacity, the output voltage is also made sensibly independent of the supply voltage. However, true voltage references – cells which generate a voltage to within very close tolerances relative to the *Standard Volt* – are rarely needed in complete systems. Their use in many cases is redundant, since there is no *measurement* of voltages, the only process that inescapably demands a reference standard. Exceptions include ADCs and DACs (although these often support the use of a common system reference voltage) and in volt-scaled components, such as the denominator of an analog multiplier, the gain-scaling of a VGA, and the slope and intercept calibration of logarithmic amplifiers used in power measurement. (In the latter case, the amp actually measures voltages, *not power* directly.) Notice that these are all nonlinear circuits. But even in systems where such components are used, it is often possible, and certainly preferable, to arrange for the use of a single voltage to scale them, either in pairs (as was shown in Section 2.4.1) or more broadly. This design philosophy, based on the dependence on *ratios,* not absolutes, can be viewed as an extension of the principles of analog design within an monolithic context, which are founded on the assured expectation of matching like against like and essentially isothermal operation.

The generation of a reference voltage to high absolute accuracy within the confines of a monolithic design, and without using trimming of any kind, involves different considerations to those previously discussed. It is no longer amenable to clever use of ratios, since *voltage is dimensional.* Circuits operating

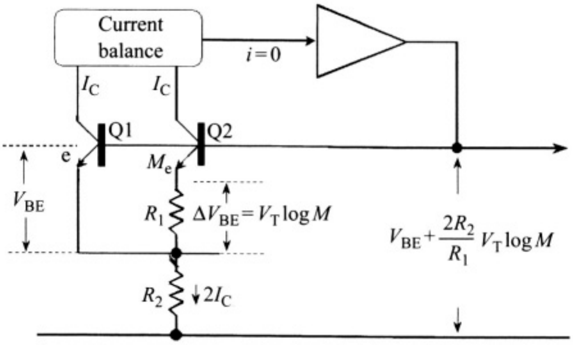

Figure 2.17. The basic Brokaw cell.

on supplies of 5 V and below will use some embodiment of the band-gap principle, such as the Brokaw cell shown in Figure 2.17. Several ways exist to get *quite close* to the intrinsic band-gap voltage of silicon, but all techniques used to realize a band-gap voltage reference are prone to fundamental sources of error of the DAP variety.

The output voltage of a band-gap reference is the sum of two voltages,[25] in proportion roughly 65% V_{BE} (sometimes called CTAT– that is, complementary to absolute temperature) and 35% ΔV_{BE} (PTAT) when using typical current densities. The latter can be generated to very high accuracy, being scaled predominantly by kT/q (a fundamental voltage) and the dimensionless logarithm of a current-density ratio $M = J_1/J_2$. This *pure ratio* can be generated in a monolithic IC to arbitrary accuracy, using unit-replicated devices and careful layout techniques, including flanking dummy elements and common-centroid placement.

It is easy to show that the sensitivity of the PTAT voltage to the value of *M* is lowered by the factor log *M*. This immediately suggests that the largest possible value of *M* should be used. There are other reasons for this choice. The wide-band noise associated with the ΔV_{BE}, due both to shot-noise mechanisms and junction resistances (notably $r_{bb'}$) is fairly large. Here comes another trade-off: the minimization of voltage noise in these cells dictates the use of high collector currents and correspondingly low resistances. This fact is non-negotiable; references operating at low currents will be inherently noisy, often dictating the use of an off-chip capacitor to reduce the noise bandwidth. Thus, when each transistor in a typical ΔV_{BE} pair is operating at $100\,\mu A$, the total voltage noise spectral density due to shot noise in the basic ΔV_{BE} is $2.07\,nV/\sqrt{Hz}$ at

[25] Sometimes the summation is performed in current mode, but the underlying principles are the same.

$T = 300$ K. Assuming the $r_{bb'}$ of the small transistor is $400\,\Omega$, this contributes a further $2.58\,\text{nV}/\sqrt{\text{Hz}}$ (the ohmic noise of the larger transistor is invariably negligible). The total noise is thus $3.3\,\text{nV}/\sqrt{\text{Hz}}$. For the commonly-used ratio of $M = 8$, the ΔV_{BE} is theoretically 53.75 mV at 300 K; this needs to be multiplied by about 9 to generate the required PTAT component (say, 480 mV) of the output, resulting in an amplified noise contribution of about $30\,\text{nV}/\sqrt{\text{Hz}}$. However, using $M = 100$, the ΔV_{BE} is theoretically 119 mV, and needs to be multiplied by only 4, resulting in less than half that noise, $13.2\,\text{nV}/\sqrt{\text{Hz}}$.

Notwithstanding these clear advantages, many contemporary band-gap designs continue to use M = 8, first popularized by Widlar and later used by Brokaw.[26] For this case, a 10% uncertainty in the emitter area ratio (which is not unlikely in a modern process using sub-micron emitter widths) is reduced by the factor log (8) to a 4.8% uncertainty in the PTAT voltage, nearly 2% of the total voltage. In modern practice, a value as high as 100 can often be used without excessive consumption of chip area. The same 10% ratio error is then reduced by log (100) to 2.17%, or 0.87% of the sum. But this is still not the total possible error in the PTAT component. The ohmic junction resistances will introduce additional components of voltage, raising the ΔV_{BE} to

$$\Delta V_{BE} = \left\{ 1 + \frac{R_\Omega}{R_2} \frac{M-1}{M} \right\} \frac{kT}{q} \log M \tag{2.13}$$

where R_Ω is the effective ohmic resistance $r_{ee'} + r_{bb'}/\beta_{DC}$ referred to the emitter branches. (Compare equation (2.10).) For example, using $r_{ee'} = 10\,\Omega$, $r_{bb'} = 400\,\Omega$, $\beta_{DC} = 100$, R_Ω evaluates to $14\,\Omega$; when operating each transistor at $100\,\mu\text{A}$ and using $M = 8$, the ΔV_{BE} is increased by about 2.3%. Using the higher ratio of M = 100, and thus a higher value of R_2 for the same current, the ΔV_{BE} error is reduced to +1.17%; when multiplied up to represent some 40% of the final reference voltage, this amounts to an elevation in output of roughly + 0.47%.

On the other hand, the control of the V_{BE} component of the "E_G" sum is much harder, since it is fundamentally "DAP", involving several production-variable parameters, including doping concentrations in the emitter-base region (determining the Gummel number), the absolute area of the emitter window, and the absolute collector current, which depends on the absolute value of the on-chip resistors. Since these are uncorrelated variables, control of the *in situ*

[26] This choice was justified when transistor geometries were much larger, and a voltage reference cell might consume a large fraction of the total die area. For this reason, it used to be common to make this one cell serve as a master biasing generator for a multi-section signal-processing circuit. Nowadays, one can often use local bias cells, to minimize coupling via biasing lines, since they can be tiny. However, this is not advised when these bias voltages are also utilized as accurate references; see Section 4.1.

V_{BE} (i.e. the operational value in the full circuit context) may be poor. For example, if we assume a ±25% variance in Gummel number (a reflection of the doping control), and a similar variation in an emitter of width $0.6\,\mu\mathrm{m}$ (the length will generally be well controlled), and further assume that the resistors (which set all the transistor current densities) also have an absolute tolerance of ±25%, the *in situ* V_{BE} might vary by as much as ±15mV, amounting to a contribution of ±1.25% in the typical output of about 1.2V. Combined with the ±0.87% random uncertainty in the (uncorrelated) PTAT voltage, and the additional systematic elevation due to R_{Ω}, the worst-case error can easily amount to –2/+2.5%.

With these various trade-offs in mind, a strategy for lowering this error will now be briefly described. The chief objective has to be the improvement in the accuracy of the main V_{BE} with further reduction in ohmic errors. This clearly calls for the use of a very large Q1, in the basic cell, which could be realized by using a much wider and longer emitter, say, $3\,\mu\mathrm{m} \times 40\,\mu\mathrm{m}$ rather than $0.6\,\mu\mathrm{m} \times 5\,\mu\mathrm{m}$, perhaps having several emitter fingers to further reduce R_{Ω}. But imagine the area that would then be consumed by Q2: it would be at least eight times larger in a standard realization, and would preferably be as much as a hundred times larger! This is an inefficient trade-off, though technically satisfactory, except perhaps with the added concern that the cell may need a larger HF stabilization capacitor (not shown in Figure 2.17).

A better approach is to separate out the cell fragment that generates the PTAT voltage and add an *independent section optimized strictly for providing a very accurate* V_{BE}, as shown in Figure 2.18. This topology is only an example of the numerous ways in which this idea can be realized, and is used here simply to make a point. An experienced designer of reference cells will be able to find several shortcomings in this sketch, and we could easily spend the rest of this chapter discussing trade-offs to improve the supply rejection, the inclusion of holistic compensation for a variety of special applications, enable-disable functions, etc.

The main principle here is that by separating the PTAT generator from the V_{BE}-determining device and focusing on optimizing the latter for minimum sensitivity to production variances, a more robust overall solution is reached. In this case, the trade-off is one of accuracy versus complexity (and thus chip area), a trade-off frequently invoked in monolithic design. Having taken that step, we might seek to extract further performance improvements out of these extra components. As previously noted, absolute voltage references are needed less often than might be thought in well-designed systems. The sharing of less accurate references across system boundaries is one of the best ways to avoid the need for traceability to an external standard. In BJT-based design based, the more common requirement is for bias currents that are PTAT, rather than “ZTAT”; these can be generated with excellent accuracy, since they

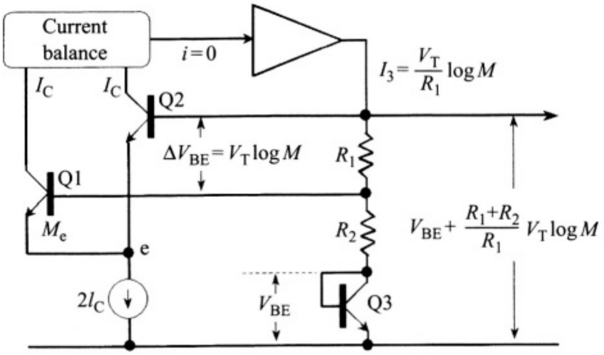
Current balance
$i=0$
I_C
I_C
Q2
$I_3 = \frac{V_T}{R_1} \log M$
$\Delta V_{BE} = V_T \log M$
R_1
Q1
M_e
R_2
$V_{BE} + \frac{R_1+R_2}{R_1} V_T \log M$
e
$2I_C$
V_{BE}
Q3

Figure 2.18. **Band-gap reference using a separate transistor for the V_{BE}.**

do not depend on V_{BE}, but solely on the logarithm of a simple ratio, scaled by kT/q.

2.4.6. The Cost of Robustness

Since robust design has so many benefits in high-volume production, with the expectation of net productivity gains through its use, it may seem odd to speak of a *cost* of robustness. What is that cost? It often takes the form of *reduced performance.* This happens because there is a kind of exclusion principle at work. One can push performance specifications aggressively, to the limits of the norms for some IC process. Assume an ultra-low input offset voltage is one of the target specifications for a competitive op-amp. Being “TAP”, such improvements have been happening for decades. Eventually a point is reached when the sheer force of process statistics stands in the way of further progress and one must pay the price. In this simple example, it may be a trade-off between tightening the test specifications and thus discarding a higher fraction of the product; alternatively, the limit values can be relaxed with the risk of being less competitive in sheer performance, but with better yields. Here, the trade-off is more the nature of a business decision, but these issues cannot be divorced from technical considerations in a commercial context.

In another scenario, suppose we aggressively extend the bandwidth of our new op-amp, to provide a more competitive product. This is more hazardous, because dimensional attributes (such as the characteristic time-constants of the higher-order poles in the open-loop gain function) vary greatly. We are betting on DAPs, which are never a sure thing. This raises the risk of the amplifier going unstable with reactive loads, risking one’s reputation for providing reliable solutions and putting a new burden on applications support engineers. The prudent trade-off in such cases is to recognize that, in volume production, one cannot afford to indulge in brinkmanship, or pursue optimistic objectives and delineate specifications which *seem* reproducible but for which no certain foundation can be provided.

One might argue here that we are confusing robustness, which is about the reduction of circuit sensitivities to process variations, with the choice of test limits that define the specifications, which is about statistics. Certainly, there is a good deal of overlap in this area. We may apply extensive testing to the design during the later stages of product development, for example, through the use of Monte-Carlo simulations, or wait until measured data from several production lots have been accumulated, to determine specification limits consistent with certain yield requirements. The former approach is limited inasmuch as fully realistic process statistics are often unavailable, particularly for a new and aggressively-scaled technology, perhaps one developed primary for digital applications and not yet characterized well for analog use. The latter approach is very costly, and delays product introduction.

This sort of trade-off underscores the great importance in choosing one's technology, system architecture, circuit topologies, signal levels and bias points with great care, and emphasizing those approaches that are inherently robust while studiously avoiding those that may be relying too much on "everything being right". As designers, it is our job to *create solutions* in which the yield/specification trade-off is tractable and definitive, rather than in need of statistical studies or the fabrication of many production lots to demonstrate.

It is in this arena that one's contribution to robustness can be most effective. Using the *same* technology, and the *same* production standards, some designers consistently achieve better yields than others. Might it be that their high-yielding parts are specified less aggressively? Probably not. A review of many designs over a period of decades shows that robustness is not a matter of slackening down to more conservative specifications. That is, "fake robustness" and would not be competitive. Rather, it is because good designers use their medium, the *tabula rasa* of the raw silicon wafer, very thoughtfully, and extract *genuine* performance advantages that have eluded competitors, while still maintaining excellent yields.

2.5. Toward Design Mastery

Each of us has a unique and idiosyncratic approach to the task of designing IC products. We acquire this personal style over a long period of time, spent in learning-by-doing, invariably by going down many dead-ends before finding the way forward, and always learning as much through our mistakes as from our successes.

At the technical level, the design of integrated circuits for manufacture is not in any fundamental way different from design in a student context. The emphasis on commercial success does not require that skills learned in an academic course of study, or in early industrial experience, be totally supplanted. But it does demand *a change of outlook,* from one in which intriguing technical challenges are the focal point to one in which these are seen as *only one aspect* of a much broader range of issues that will consume a large fraction of the available time. *Circuits are not products.*

Design for manufacture means that the professional needs to constantly keep in mind the singular, long-term objective of either *satisfying an existing customer demand* or *anticipating an unarticulated need* and providing a ready solution. In the best outworking of the latter scenario, one can literally *create a market* for innovative products, when these address a problem that was not obvious until the solution was offered. Product design requires a compelling, consistent and unrelenting vision of the end-game. It demands a candid and auto-critical view of all of the numerous ways in which the project can fail. The technical aspects of this challenge are very significant, perhaps even dominant, but in a commercial context the circuit design phase must be regarded as but one contribution to the success of the overall product development.

2.5.1. First, the Finale

Maxim: Product development starts with the objectives, not the availables.

This simply means that the starting point for any well-run IC project is a total comprehension of the proposed product, addressing a real need as component part of a business-development strategy. It must entail a clear understanding of what *will* be achieved in the course of the development; the competitive (and often novel) attributes which it *will* possess when delivered to the customer, at a certain time and at a certain cost that *are already* determined; the performance specifications that *will* be met at that time; the package that *will* be used; the testing methods that *will* be used to ensure performance; and similar aspects of the outcome. This is very unlikely to happen unless all of the objectives and the schedule have been agreed to by the team and the needed resources have been identified and assigned in advance.[27] A common precursor of the development is the preparation of the product definition document.

The alternative stratagem, *starting with the availables,* means that someone has a "promising new idea for a circuit", and an unscheduled project begins right away to embellish this idea in a product. The strategic value of the product has not been ascertained, nor are the objectives clear. The project very

[27] Since at any given time a corporation is bounded by finite resources, the addition of a new project unavoidably means that fewer resources will be available to handle a large portfolio of existing projects. In a well-run organization, the impact of new projects on existing ones can be automatically accounted for by sophisticated project management and scheduling software.

probably arises in isolation and may proceed without an awareness that similar (possibly more successful) work is being pursued elsewhere in the company. Interestingly, maverick projects that have this sort of genesis are *not necessarily destined to failure.* They may actually turn out to be tremendously valuable, when eventually converted into an *outcome-oriented* project, perhaps needing significant changes in the design.

2.5.2. Consider All Deliverables

Maxim: All of the project deliverables should be identified right from the start.

These may be divided into *external* and *internal* deliverables, and are all the things that must be generated for delivery either to the customer (externals) or to development/manufacturing (internals), at various times between project start and the Product Release date.

Examples of external deliverables[28] include:

- The *Data-Sheet* – essentially a contract between the supplier and the customer.
- *Product samples,* packaged (or known-good die), tested to Data-Sheet specifications in the quantities needed to satisfy anticipated evaluation demands.
- *Application Notes* for standard catalog components, which elucidate the many ways in which the product can be used, through very specific, fully-worked examples.
- *Evaluation Boards* for high-speed and special components.
- *Reference Designs* for such things as a communications chip-set.
- *Software, Firmware and Development Systems* (for digital ICs).

Examples of internal deliverables include:

- *Detailed Product Specifications,* for use throughout the development, and defining many internal sub-objectives; usually a super-set of the Data-Sheet.
- *Project Schedule and Plan,* delineating the major milestones (Concept Review, Design Review, Layout Review, Wafer Starts, First Silicon, Evaluation Completion, First Customer Samples, Product Release, etc.) and identifying needed resources.

[28] These will generally be needed for internal development purposes, also.

- The *Product Description Document,* which should be generated as an accumulative body of material, and will include such things as marketing and cost data, overall system and circuit theory, block diagrams, cell schematics and detailed descriptions of circuit operation, results from simulation studies, test methods, usage schematics, application ideas, etc. The responsibility for generating this important internal document will usually be shared amongst several people, all of whom need to be advised as to what is expected of them in this regard.
- *Wafer-fab Documentation,* including process type, manufacturing site, lot sizes, etc.
- *Assembly Documentation,* including package type, die attach method, bonding diagrams, use of over-coats, etc.
- *Test Documentation,* including the complete delineation of the tests needed at wafer probe, full descriptions of the support hardware, details of trimming algorithms (where used), and similar details for final test, including all limit parameters.
- *Reliability Documentation,* including life-test and ESD results, production quality monitoring, failure analysis, outgoing inspection, etc.

It is unrealistic to expect that all elements of this large body of information will be available at the start of a development. However, the basic philosophy here advocated is that a very comprehensive plan must be on record before significant design resources are invested, with the certain expectation that the documentation will expand as the project proceeds. This perspective is clearly quite different from the notion of starting with a brilliant circuit concept and immediately proceeding to develop it, in the hope of it becoming a product.

2.5.3. Design Compression

Maxim: Complete the basic design within the first few weeks of the project.

One of the easiest traps to fall into when undertaking a product development is to assume that the available time, delineated in a master schedule, will be spent in a fairly homogenous fashion, being a sequence of design studies and associated simulation experiments or verifications, occurring in a steady, constant density throughout the project. However, experience teaches that *very considerable time* must be allowed for all manner of work related to validation and presentation of the design, in preparation for a Design Review and transfer to mask layout, even when the “design” is well advanced.

For example, suppose one has assessed the need for a 12-week design period, and formally agreed to this schedule. Bearing in mind the maxim “First, the Finale”, it can safely be assumed that the material needed for presentation at the

Design Review should be delivered for peer consideration at least one business week prior to the date set for that review. This material minimally consists of the following: A complete set of well-annotated schematics (clearly showing all device sizes and special layout notes, bias currents at the top and bottom of each branch, internal voltages, high-current branches, etc.); a comprehensive collection of simulation results (the good, the bad and the ugly: i.e. worst-case performance, for process, supply voltage and temperature corners, and with mismatch effects, rather than just the nominal results); and a text that puts the product into perspective, outlines any necessary theory and provides a component-by-component description of circuit operation, illustrated with more basic figures than the detailed schematics. Such a document is likely to take *at least* a week to prepare, and probably longer.

This suggests that one can expect to lose between one and four weeks at the end of the nominal design period. Prior to such "wrap-up" work, time must be allowed for numerous simulation studies to be performed on the *complete product,* even if the need for this has been minimized by careful attention to cell boundaries and through rigorous verification of these smaller entities. Analog cell interactions are common, whether through bias or supply lines, or subtle substrate coupling effects; some may be serious enough to warrant a significant change in overall structure. For a complex product, these top level simulations will be quite slow and time-consuming. In this connection, it is prudent to include all the ESD devices from the very start (one sometimes needs to devise special, pin-specific ESD protection schemes), and be sure to use a complete model of the package impedances and the mutual coupling between bond-wires. Keep in mind *that fast transistors are not aware of your expectations.* Given the slightest excuse to burst into song, they will.

When such time-sinks are anticipated and identified, a basic rule becomes apparent: the nominal design should be completed with a very short span of time, a matter of a few weeks, right at the start of the project, rather than allowed to gradually evolve over *the full length of time* scheduled for it. This overarching objective can be facilitated by adopting a sort of "imagineering" approach, in which the first item to be entered into the schematic capture domain is the *top level schematic,* which should be drawn as a pseudo-layout (e.g. see Figure 2.19). At this stage, it is acceptable to simply draw cosmetic boundaries for the main sections, whose sizes are estimated only approximately, knowing their general contents.

This layout-style schematic will show all the bond-pads to scale, the ESD protection devices and their power-busses, and allows one to connect up the blocks using actual "wires", provided that the cells are assigned pin symbols. Inside these temporary blocks can be ideal elements, such as independent and dependent sources, chosen to crudely represent the block's function, or perhaps some previously-developed cells. When this is completed, the top-level

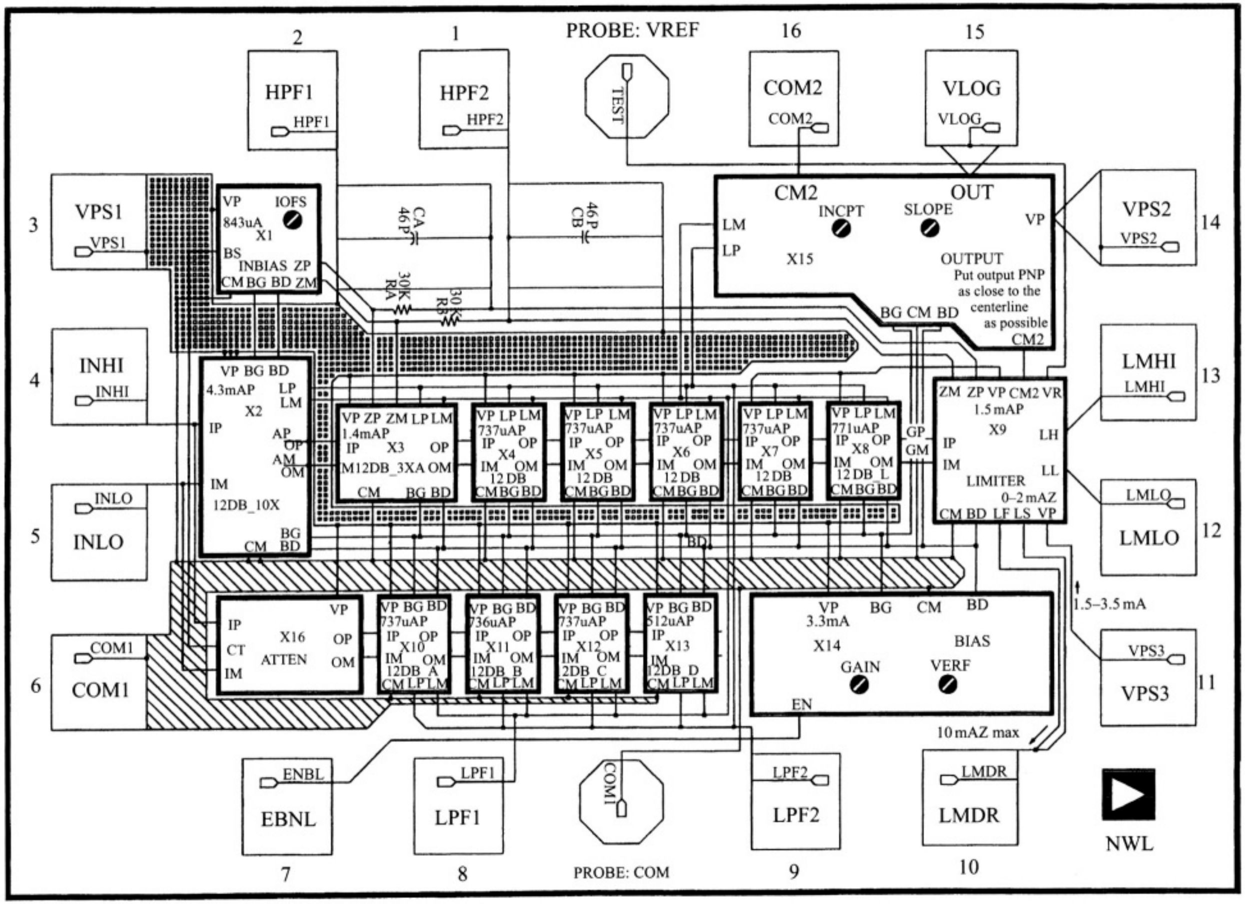

Figure 2.19. A schematic that is also an IC layout.

schematic should be error-free when generating a net list. One can thus have a *complete product schema by the end of the first day or two.* During the next few weeks, these blocks will be progressively fleshed out as real circuits, though still permissibly using some ideal elements, starting with the blocks most likely to prove challenging and needing the most invention. Although crucially important, the design of bias generators can usually be deferred until later in the project, although there may be exceptions as to any rule.

To readers unfamiliar with this approach to IC design, it may sound hopelessly idealistic and not the sort of thing one can really implement. However, the author has been using exactly this method for many years, and it is not merely workable, but very effective and time-efficient. It forces one to pay attention to the objectives – the finale – from the very start. It requires a full consideration of the pad sequence and the optimal location around the chip boundary. This in turn leads to a well-planned "street plan", showing the most important routes (such as those for the primary signals and the power supplies) and every one of the less critical, but nonetheless necessary, auxiliary connections, for biasing and control purposes. It invites one to use whatever means available to clearly indicate which of these routes must have a high current capacity or an especially low resistance (for example, by widening a "wire" into a narrow rectangle, and using cross-hatching to make these major routes very clear); or which must be extremely short or narrow, to minimize parasitic capacitances; or which paths must be kept apart to minimize coupling, or made equal in length for delay balancing, etc.

Special treatments of this sort are going to need articulation sooner or later, and to the extent that many such details can be foreseen and dealt with very early in the project, they are best got out of the way before the more troublesome mannerisms of the juvenile product begin to appear. The method is also a fine way to feel a strong sense of progress toward one's goals and to add a palpable reality to the development, on which stable platform the design can proceed with greater confidence. The alternative is to nibble at matters of cell design for weeks on end, a little here and a little there, with the hope that everything will fit together in the end; this is the antithesis of design mastery.

2.5.4. Fundamentals before Finesse

Maxim: Emphasize the use of strong basic forms; use clever tricks sparsely.

A study of a large cross-section of IC designs would almost certainly show that the ones that gave the least trouble in manufacturing were those which used strong, elegant techniques, often involving a minimal number of components, and appealing to holistic principles, in which one cell enters into a close and comfortable synergy with its surroundings. Conversely, products which are difficult to manufacture are invariably found to appeal to a lot of "super-structure" to fix up one source of error after another, or address performance short-fall. In an actual limited study, looking for root-cause-of-failure in about two dozen products, and conducted about 17 years ago, the Pareto analysis revealed that "Design Methodology" was responsible for nearly 30% of all failures in first silicon. Adding in those failures due to "Difficulties in Simulation" and "ESD Protection" brought this up to 72%. The remainder of the failures could be traced to inadequate modeling accuracy, layout errors, omission of interconnect parasitics, and various errors in the schematics. In this particular study, none of the failures were due to manufacturing mistakes.

Although a limited and dated result, it does point to the importance of attending to *the fundamentals of design.* Some questions to ask at frequent intervals are: Is this component (in a cell) essential? When the Design Document is written, how will I justify its inclusion? What would be the impact on performance if it should be removed? Not all components should be excluded just because they play a minor role. Their combined contribution to robustness may be valuable. But in thinking at every turn about the *purpose* of adding one or more components to an otherwise satisfactory design, one can reduce the risk of unwittingly introducing future and possibly time-consuming problems.

2.5.5. Re-Utilization of Proven Cells

Maxim: Do not re-invent the wheel; adapt the trusted form.

This is actually a surprisingly hard lesson to learn. Those of us who enjoy cell innovation spend a lot of time thinking about alternative ways to achieve certain aspects of performance that have already been met numerous times before. Such activity is not to be discouraged; it is the well-spring of important new ideas, and may be considered an appropriate response to the previous maxim, reworded as: *Always be on the look-out for new fundamental forms.* Nevertheless, time-to-market pressures require that we re-utilize existing cells whenever possible. The savings in time, and *a potential* reduction of risk, come from several sources:

- The needed cell design (or something close) is already in hand.
- It will often be proven and de-bugged; a body of performance and test data for actual material will be available.
- The cell layout also already exists; while this may undergo some alterations in the new context, the general form of this layout and its subtleties can be preserved.
- Re-use eliminates time-wastage in chasing newly-invented bugs.

On the other hand, there are several reasons why cell re-use is not quite so easy:

- The needed performance will invariably differ (from slightly to radically) to that provided by the available cells, requiring varying degrees of redesign.
- The descriptive support of the cell may be minimal or even non-existent.
- The adoption of someone else's cell design without fully understanding it, and the *context* within which it was developed, can be hazardous. For example, taken in abstraction, an available voltage reference cell may appear to perform well, and the schematic annotation to the effect that "$Z_{OUT} < 1\,\Omega$" seems reassuring. However, the original usage of the cell did not require a low output impedance at 100 MHz, as does your application, and Z_{OUT} was actually measured at 10 kHz, although this was never noted. Without a meticulous assessment of its *suitability* to the present environment, this cell could contain the seeds of problems further down-stream.
- The available design may be on a different process technology to the one needed for the current project.

Closer consideration shows that there are really *two types of re-utilization.* The one that is generally discussed involves the adoption of someone else's work from a library of cells, found in an internal memorandum or company web page, presented at a design review, or by familiarity with the work of a team member. But an equally important class of re-utilization is that based on the proven concepts and cells that a designer carries around in his or her head. Skillful re-use of ideas, the essence of experience, is usually a far better basis for robust design than the opportunistic adaptation of somebody else's work.

2.5.6. Try to Break Your Circuits

Maxim: Don't pamper your circuits; make them confess their darkest secrets.

Designers enter into a kind of love affair with their circuits. Sometimes, this takes on a parental aspect, and due attention is paid to making sure that discipline is administered when needed. We are usually quite thorough in putting our progeny through a series of challenging experiences in readiness for the harsh realities of the world beyond the workstation screen. But there is also a curious inclination to be kind and considerate: we may avoid subjecting the design to more than it can bear. Such compassion for a circuit cell is unwise. The world of real applications will certainly not give your product an easy ride: neither should you.

An important function of the designer is to routinely and relentlessly push a cell design to the brink of disaster and then bring it back again to the placid waters of normal operation. “Routinely” in this connection means at least several times a day, from the earliest moments all the way through to pre-layout final checks. “Relentlessly” means with no concern for the possibility that the design will break under stress. Such attempts to break the cell, or reveal its secrets or some hidden pathology, will include the use of numerous *parametric sweeps.* Most modern simulators allow a wide range of interactive sweep modes, in which any desired parameter can be identified and swept over massive ranges. Some of the more obvious:

- *Supply voltage*: if the nominal supply is 2.7–3.3 V, sweep it from 0 to 10V. You do not expect the circuit to work at zero, nor do you expect it to collapse in a heap at 10 V (though there may circumstances when this would be an unreasonable stress). Do this using both DC and time-domain sweeps. Use sweep-from-zero *and* sweep-to-zero exercises: these will tell you a lot about start-up and minimum supply limits. Perform these exaggerated supply sweeps at very high and very low temperatures, and using process corner models.
- *Temperature:* if the normal operating range is –35°C to +85°C, that should not prevent you *wondering* about what happens at –75°C or +175°C (the workstation will not melt) perhaps while using supply voltages at least 20% bigger or smaller than the nominal range. Frequently, one will observe several anomalies at temperature extremes. For example, the gain of an amplifier that is supposed to be 20 dB may show a sudden drop above 115°C. Since this is well above the required operating temperature, it could be ignored. Nonetheless, good design practice requires that one *immediately* picks up this trail and finds the root cause, even though a remedy may not be implemented. The discovery of all such pathologies revealed by swept-parameter experiments should be treated in this way. In many, many cases, these digressions lead to valuable new insights, and reveal incipient weaknesses that could threaten yields or result in field failures, when combined with some unhappy combination of supply voltage, temperature, process corners and device mismatches.
- *Do not stop there:* sweep *everything!* For example, sweep all sheet resistances from one half to twice their nominal value; BJT betas from a one-third to at least five times their nominal value; and so on. If it is found that the performance aspects that ought to be TAPs are not, one must ask why.

2.5.7. Use Corner Modeling Judiciously

Maxim: While “Corner Models” are often more myth and guesswork than definitive, put your prejudices aside and use them anyway: they can be most revealing.

The use of so-called Corner Models is somewhat unfocused. These models are generated by the team producing device characterization data for simulation purposes, and they invariably involve a certain amount of guesswork. For example, in a pure-bipolar process, the transistor models for one extreme may simultaneously (1) maximize all the junction resistances, including R_E, R_B, R_{BM} and R_C; (2) maximize all the junction capacitances, including C_{JE}, C_{JC} and C_{JS}; (3) minimize the saturation current I_S; (4) minimize the DC beta parameters, including BF; (5) maximize the transit time T_F, and so on. (The total number of parameters is more than 40 in the full set for a BJT, and most are treated in a similar fashion.) These extreme values give rise to what may be called the "SLOW" model, as a little consideration of the effects of these changes on circuit performance will show.

In addition, the "SLOW" library will set all resistors of every type to their maximum value, by using the maximum sheet resistance, the most extreme reduction in resistor width and the most extreme extension of resistor length. It will likewise set all the passive capacitors at their maximum value, by assuming the minimum oxide thickness and the largest expansion of the area. In some cases, this rigour will include the wiring parasitics, using a similar set of considerations. Similarly, other components, such as ESD and Schottky diodes and inductors available in the process are pushed to their "SLOW" corner. Of course, the "FAST" models reverse this process. The treatment of corners in a CMOS process is essentially the same, with similar objectives, although greater effort is expended to include the correlations between electrical parameters, based on a smaller set of physical parameters.

Note that, in using corner models, the designer is left determine the extreme temperatures and supply-voltage conditions which result in the most severe degradation in performance. A full matrix of results, for just one aspect of performance, requires no less than twenty-seven simulation runs: One uses the "SLOW", "NOMINAL" and "FAST" models, and in each case the minimum, nominal and maximum temperatures (say, – 60°C, +30°C, +130°C, even though actual operation may be limited to a smaller range – in the spirit of trying to "Break The Circuit", or at least exploring where it begins to sweat); these are repeated for the minimum, nominal and maximum supply voltage (say, 2.6, 3 and 6V). A convenient way to view these results is by using a set of three pages, one for each supply voltage, each comprising three panels, one panel for each model parameter set, and each of these having the swept parameter along the horizontal axis, and the three temperatures in each panel. If this is to be repeated for each of the critical parameters (which ought to correspond closely to the line items in the data sheet, where possible), hundreds of pages of results may be needed to fully capture the full (PVT) corner performance, and all these experiments will invariably (but unwisely) presume that matching remains perfect.

In practice, the use of these corner models is quite problematical, for several reasons. To begin with, they quite clearly represent a very extreme state of affairs, unlikely to occur *simultaneously* in practice, or even as individual extrema. Since "worst-case" values are assigned, these are presumably the limit values for which a production wafer would actually be *rejected.*[29] So, the first problem is whether to believe they are at all *realistic.* The second problem with corner testing is that it really does not show the worst case that might arise, when mismatches are included. Indeed, an otherwise flawlessly robust circuit might continue to work very well at the corners, *when all the devices match,* then collapse seriously into a mere shadow of its former self when realistic mismatches are included. Third, it may happen that local performance minima actually arise *somewhere inside* one of the ranges of worst-case extrema, which are not captured in corner studies. Or, it can arise from *a combination* of some unfortunately set of parameter values *and* certain mismatches. Fourth, these studies do not provide much insight, if any at all; they simple demonstrate a lack of robustness, without clearly pointing the way forward. Finally, it will be apparent that a huge amount of time will be needed to provide a comprehensive set of corner results. Regrettably, even a small change to the design may necessitate repeating these tedious procedures.

What we have here is a *most fundamental kind of trade-off*: that between time-to-market and risk. The use of comprehensive corner testing is inefficient. The objective of any product development is to first, exercise dominance over the material, and *dictate what the circuit shall be permitted to do,* rather than treating the challenge as something like science, which is the exploration of a domain not of one's own making, to try to understand its inner mysteries. The true purpose of one's studies throughout the design phase should be *the minimization of enigma and the maximization of insight.* As already stated, these objectives are best tackled by the routine use of *sensitivity* studies at every point in the design. If one minimizes all the major sensitivities independently, there is a high probability that the overall system will be inherently robust. Then, when small changes are made, one can be fairly sure about the consequences, and the need for time-consuming re-runs of the corners is minimized.

While these cautions are based on reasonable enough concerns, there is at least one reason why the use of corners may nonetheless be of benefit, and it is a little subtle. It was noted above that the algorithms built into the corner modeling include variations in resistor width (and other similar narrow dimensions). It has also been noted that component mismatches can destroy circuit integrity,

[29] These are based on measurements made on production-specific test sites, which are often placed at just five locations on the wafer, but sometimes embedded in the scribe lane between the chip boundaries.

and that to mitigate against these, one should routinely use equally-sized unit elements when building up large ratios or striving to maintain an exact equality of component value. Now, depending on one's schematic capture software, and the way in which these structures are defined, it is possible for errors to arise in the way the software interprets device scaling data. In turn, this may either reflect badly on the performance, or it can hide sensitivities.

For example, in that amplifier we developed (Figure 2.5), three $3\,\mathrm{k}\Omega$ resistors were used in parallel to generate a $1\,\mathrm{k}\Omega$ component, and four of these same units were connected in series to generate a $12\,\mathrm{k}\Omega$ component. Suppose one first decided to make each resistor $5\,\mu\mathrm{m}$ wide and $15\,\mu\mathrm{m}$ long, when the sheet resistance is $1\,\mathrm{k}\Omega/\mathrm{square}$. Then, in the schematic capture environment, the $3\,\mathrm{k}\Omega$ element might be denoted as a single resistor with a length of 15 ("microns" being assumed by the program) and a width of 3∗5, the multiplier being necessary to satisfy the subsequent verification of the layout against the schematic. Likewise, we might denote the $12\,\mathrm{k}\Omega$ element as a single resistor with a length of 4∗15 and a width of 5. These will automatically be calculated in the net-lister and the simulation results will be correct.

However, the width of 3∗5 may be treated as 15 and the length of 4∗15 may be treated as 60; information about *structure* is thereby lost. The layout verification software will be happy, because it is told to measure the *total width* (for the $3\,\mathrm{k}\Omega$ resistor) and the *total length* (for the $12\,\mathrm{k}\Omega$ resistor). But now we apply the corner models, let's say, the SLOW models. With this representation, the $1\,\mathrm{k}\Omega$ resistor width is reduced by only one delta-width unit on each side (just by way of example, say, $-0.2\,\mu\mathrm{m}$), and the $15\,\mu\mathrm{m}$ becomes $14.6\,\mu\mathrm{m}$, while its length is increased by only one delta-length unit at each end (say, by $+0.35\,\mu\mathrm{m}$), to $15.7\,\mu\mathrm{m}$. Its apparent "worst-case" value (neglecting the sheet resistance, which affect all units equally) is thus $1.0753\,\mathrm{k}\Omega = 15.7\,\mu\mathrm{m}/14.6\,\mu\mathrm{m}$. Working through similar arithmetic for the $12\,\mathrm{k}\Omega$ resistor, it has an apparent value of $13.196\,\mathrm{k}\Omega = 60.7\,\mu\mathrm{m}/4.6\,\mu\mathrm{m}$. So the ratio is no longer 12, but 12.27. This is likely to be a "false positive": the use of strict unit elements will *guarantee* this ratio in the presence of any width and length variations.

Counterexamples arise in which less than careful attention to this sort of possibility will *obscure* real sensitivities. It should be added that not all schematic-capture software will suffer from this particular source of error. When in doubt, the safest approach is to explicitly include all of the units in such an ensemble, even if the page gets a little cluttered, or relegate them to a sub-circuit. An secondary advantage of the explicit approach is that it forces one to remain *fully aware of the physical reality* of one's circuit, and to remain focused on the constraints of layout. For example, resistor dimensions may snap to $0.1\,\mu\mathrm{m}$ increments, so avoid the use of ohmic values in very precise applications, and state this value in terms of length and width. allowing the netlister's knowledge of sheet resistance calculate the ohms.

Always keep in mind that there is never a worst case in the on-going production statistics for a product. There are good cases and there are bad cases. The art of design is to ensure that there are far more of the former than the latter.

2.5.8. Use Large-Signal Time-Domain Methods

Maxim: Do not trust small-signal simulations; always check responses to fast edges.

Elaborate use of, and an excessive reliance on, Bode plots and other small-signal methods is extremely risky. One might use these initially, and briefly, to generally position the AC behavior of a circuit, and occasionally as the design progresses, and again in generating the supporting documentation for a Design Review. But as a general rule, the circuit should be subjected to strenuous time-domain exercises during the product development. These will sometimes use small test signals (say, millivolts), during which the correspondence between the AC gain/phase results and the time-domain should be very good.

On the other hand, it is not at all uncommon for these little "tickler" signals to persuade the circuit to launch into a swell of oscillations, if it is prone to do so. This can happen even when the AC results appear to be satisfactory, but perhaps one has paid to much attention to the gain magnitude, which appears to roll off gently and benignly, with insufficient concern for the phase. Even when these really do predict a satisfactory stability margin, only slight deviations from a quiescent bias point can quickly change all that, in many classes of circuits.

When pursuing such experiments, one may also be inclined to choose a rise/fall time for the excitation that is consistent with the system requirements, say, in the 10ns range for a 10 MHz amplifier, having an intrinsic rise-time of about 35 ns. However, the circuit may exhibit some unexpected pathology when driven from very fast edges, perhaps as rapid as 10 ps. Though the circuit will never encounter such signals in practice, the lessons one can learn from ultra-wideband excitation are often unexpectedly valuable, revealing nuances in the response that call for immediate remedial action. Such investigations should be conducted over the full (even an extreme) range of temperature, and at process corners, even when the behavior under nominal conditions appears trouble-free. In this connection, it is also important to use fast excitation sources when the circuit is driven with much larger signals. Overdrive conditions may reveal yet other conditional oscillations, as devices approach saturation or their bias conditions cause a large change in device inertia.

2.5.9. Use Back-Annotation of Parasitics

Maxim: In simulations, a "wire" is just a node of zero extent. But an integrated circuit has many long wires which have capacitance to substrate and to each other. Don't neglect these.

Many of the differences that arise between the measurements made on silicon circuits and the predictions of simulation can be traced to these parasitic capacitances. One is inclined to neglect the extra rigor needed to extract these from the layout and verify performance with their reactances included, particularly when the circuit is only required to meet some modest low-frequency objectives. Clearly, when high frequencies are involved, such back annotation is mandatory. Many problems can arise from the loading of cells by the shunt capacitances to the substrate (particularly when using low-g_m CMOS, that may look fine until one adds a few femtofarads on its output); or from the coupling between these interconnects; or from mismatches in these capacitances that can affect certain aspects of circuit balance.

In speaking of capacitive coupling to "the substrate", one is bound to ask: What node is that? It is certainly not "ground", that is, the external reference plane that is customarily regarded as a node of "zero potential", identified in SPICE by the node name "0". The choice will vary from one technology to another. It may be satisfactory to use the paddle on which the circuit is mounted as that node; be aware that this will differ in potential from the external ground when the full package model is included, which should be standard practice whenever a modern high-speed technology is used – for whatever purpose. It may be necessary to divide the chip area into different zones for the purpose of defining these various "local grounds".

In a similar way, be very careful in selecting the appropriate node for the substrate connection to all devices (not only transistors, but also for the supermodels of resistors and capacitors). This should *never* be "0" in a monolithic product, and it may not always be correct to assign it the node name for the paddle. Frequently, different areas of an integrated system will need to use independent node names to identify the appropriate substrate potential for the various devices or blocks. The most accurate identification and partitioning of these important nodes can usually be determined only after reviewing a preliminary layout.

2.5.10. Make Your Intentions Clear

Maxim: Understanding every subtle detail and fine point of your masterful design is great. Now, take steps to ensure that everyone else on the team does.

We are inclined to assume that what is "obvious" and "only common sense" will be apparent with equal force to our co-workers. However, it often will not be. This is not a commentary on their intelligence, but invariably due to a lack of clarity in stating your precise intentions. One of the more critical team interfaces is between the schematics and the layout designer. If you are lucky enough to work with very experienced colleagues, you may be able to take the risk of presuming that they will do certain things just the way you would

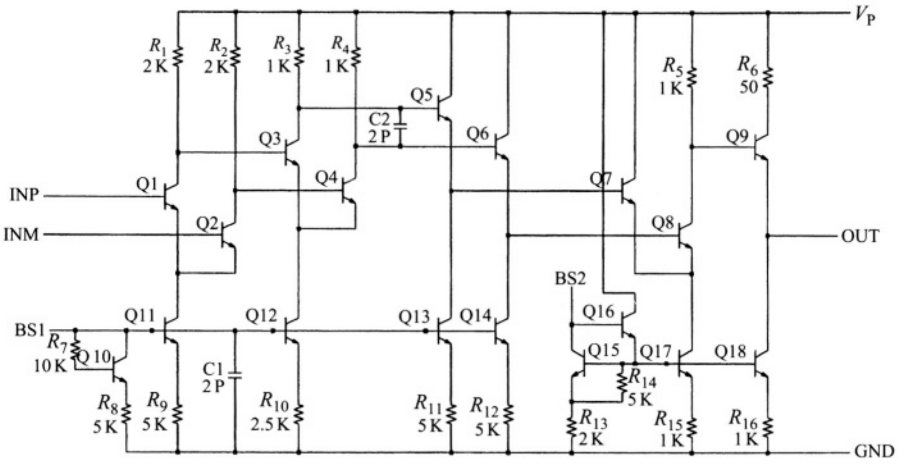

Figure 2.20. A lazily-constructed schematic having numerous ambiguities.

(i.e. the way which is absolutely critical to ensuring performance, but you did not say so).

Consider a simple example in the annotation of a schematic. Figure 2.20 shows a lazy-minded drawing of the circuit. Try writing a list of at least ten mistakes that could be made by the layout designer, acting solely on this schematic. Now examine Figure 2.21, which avoids these traps by explicitly noting certain critical requirements. Of special importance are those related to metal connections and the identification of locally merged nodes. The simulator will be quite indifferent to how the schematic is drawn in these areas: a node is just a node, having zero physical extent. But the silicon realization will be significantly impacted by a lack of attention to the use such local merging, because of the resistance of the metal traces, which will in some cases have non-local currents flowing through them. These resistances may need to be extracted from an interim layout. However, when properly indicated on the schematic and connected accordingly, and balanced in length if necessary, these small intraconnect resistances will often not matter. If nodes are allowed to be incorrectly connected one should be aware of the potential for malfunction.

2.5.11. Dubious Value of Check Lists

Maxim: Antibiotics are valuable. But it's much better to stay healthy.

Relying on check lists to achieve a robust design is hazardous. When used prior to a Design- or Layout Review, they may be of value in catching a few straggling indiscretions. Consulted religiously on a daily basis throughout the design and layout phase, they might be useful in trapping mistakes in the

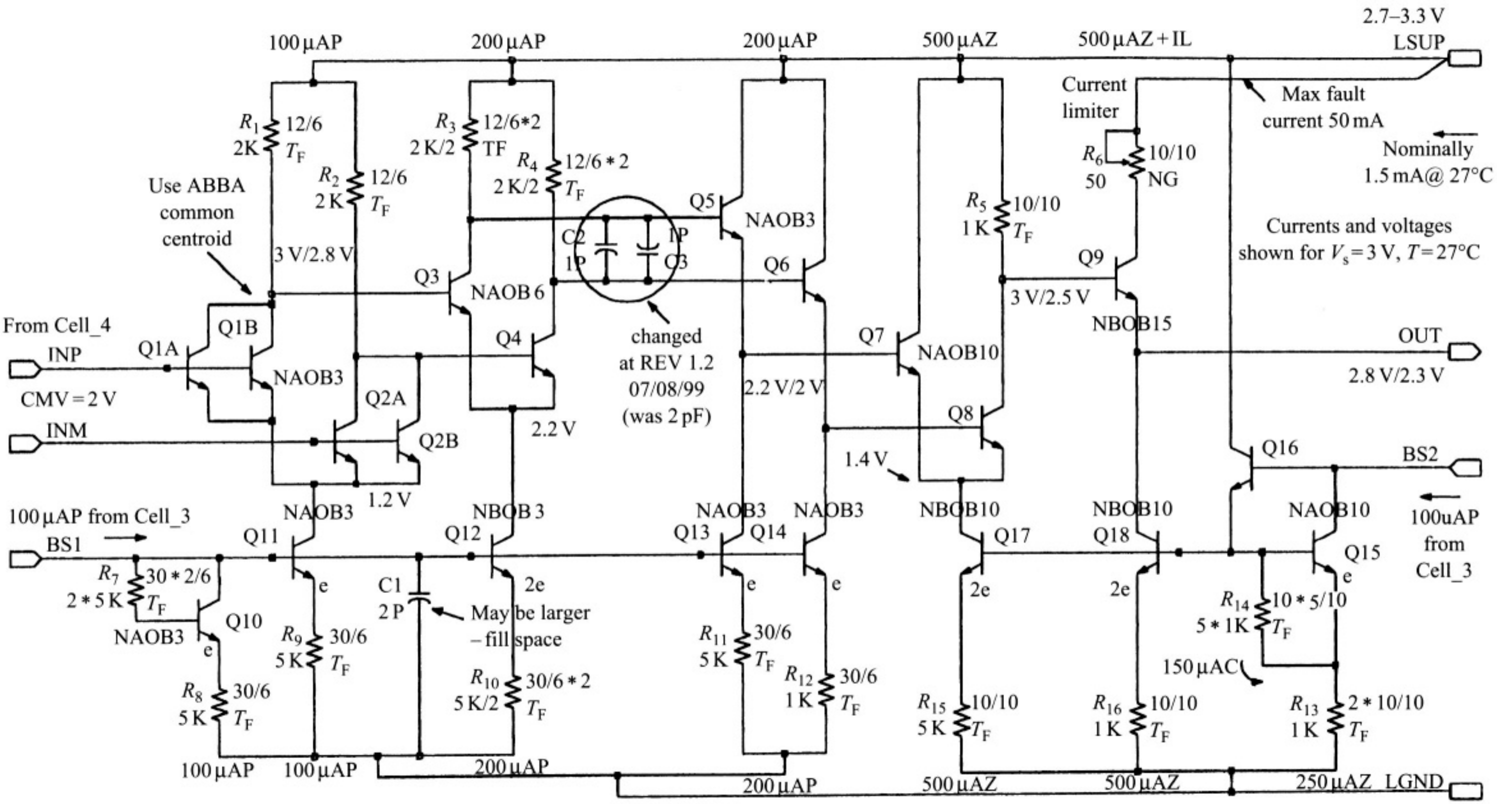

Figure 2.21. Self-documenting schematic identifying special construction needed at layout.

making. But there is a danger in either case that one may gravitate toward a mode of design that is reminiscent of painting by numbers, or responding to a multiple-choice questionnaire; that is, by reacting to a prompt for some *pre-specified* action, rather than by independently deciding what the right action should be at each juncture.

Check lists tend to be superficial, stating broad and often comically commonsensical truths. They touch on a limited set of issues, and may overlook major areas of concern. Some of the questions (such as "Did you simulate your circuit over a full range of operating conditions?") will appear downright stupid and naïve. These may prompt the person sincerely wishing to extract some value from the checking process to wonder whether to spend any further time with the rest of such rules. At the other extreme, specific operational problems that have arisen in connection with previous developments may seem too arcane to include in a general list.

However, check-lists have their place. In the pursuit of robust design, and the minimization of time-to-market, it probably does no harm to review the issues they raise, if time is available in the rush to get your product into wafer-processing. You might seek ways to add your experiences to these lists, particularly those relating to unexpected anomalies. (A well-structured system for the capture and retrieval of information is needed.) In the spirit of Total Quality Management (TQM), the check lists should continue to grow in value, particularly to new recruits, as additional non-obvious pitfalls and sources of failure become apparent.

2.5.12. Use the "Ten Things That Will Fail" Test

Maxim: After finishing the design and layout, subject your product to an end-of-term exam.

We have struggled with many challenges in getting our product this far, to the layout stage, and may understandably be disinclined to try yet more ways to break this prize design. But it is far better to discover these, if they exist, before the costs begin to escalate, and delays accumulate in wafer fabrication. So the idea here is to project one's mind forward to the time when first silicon will be available, and vigorously play a few more. What if? scenarios, in an attempt to find the skeletons in the closet. Ask such questions as "When the supplies are applied to first silicon and the currents are found to be excessive, how might that occur?". One possibility: an additional ESD diode somehow got added at the last minute, and a full re-check of the layout against the schematic was not conducted, since "this is such a trivial change". But it was wired in *reverse polarity.* Another scenario: You did a pretty good job of indicating which interconnections must be wide, or short. But have you included the resistance of the longer, unspecified traces back into the circuit? There has been much gnashing of the teeth over such “minor” details!

Attempt to draw up a list of ten such errant possibilities; then, implement stern remedies.

2.6. Conclusion

The path from concept to customer is unquestionably a tortuous one. Choices of architecture, cell structure and technology must be made. Many vexing trade-offs will have to be faced; these are in every respect human decisions based on experience and judgment, sometimes arbitrary but never algorithmic. Many errors of both omission and commission can occur in the development of an integrated-circuit product. Making the best choices about *all aspects of performance* is just the beginning of a long journey, but nonetheless the essential starting point. It is given greater substance by generating the data sheet in as complete a form and possible, leaving placeholders for all the characterization graphs that will eventually be included, and describing all the features, applications and circuit theory, as if the part really existed. This will be your anchor through the entire journey to the customer’s door.

The bulk of the design should be *compressed* into the first few weeks of the development, leaving plenty of time for validation and verification of robustness. Begin by preparing a top schematic that is a *pseudo-layout,* with all sections clearly identified, of about the correct size and positioned correctly on the floor-plan. As the inner details gradually fill in, make sure that all of the relevant details are captured in this one document, in the same spirit as in preparing a set of architectural drawings. While supporting documentation will be essential for a Design Review, for Product Engineering purposes, and as part of a permanent record, the schematics themselves should be a complete, detailed recipe for the construction of the layout, as well as a means of communication to all who need to understand the product.

The extreme sensitivity of an analog circuit to production parameters poses especially daunting challenges, in finding a suitable overall form, in realizing optimal cell topologies and in rationalizing and regulating their operation. Conflicts will need to be resolved by making compromises, deciding between many possible directions and trade-offs, minimizing every conceivable sensitivity, and much else of a circuit design nature. Furthermore, one must enter deeply into a consideration of worst-case behavior, using corner models, extreme temperatures and the limit values for supply voltage. After the basic electrical design, the most minute details of the chip layout will need your full consideration, as well as the numerous ways in which the package will impact performance, such as chip stresses, bond-wire reactances, substrate coupling over a noisy header, and much else of an highly practical nature. Thermal management is often an essential aspect of the packaging phase.

This chapter has presented a cross-section of representative trade-offs, and proposed a few methods to ensure robustness. It will be apparent that this is not by any means the whole story. The matter of substrate coupling is becoming very important, not only in mixed-signal systems on a chip, but also in pure analog and strictly digital products. The topic of designing for testability similarly needs close attention and planning.

Circuits are not products. Circuit design is but the starting point for the numerous corrections, adjustments and adaptations that will inevitably follow, accumulating increasing delays as the project rolls along, unless the author's experience is an unfortunate aberration.