ELECTRONIC ENGINEERING

ANALOG IC DESIGN with LOW-DROPOUT REGULATORS

GABRIEL ALFONSO RINCÓN-MORA

CHAPTER 1
System Considerations

1.1 Regulators in Power Management

Supplying and conditioning power are the most fundamental functions of an electrical system. A loading application, be it a cellular phone, pager, or wireless sensor node, cannot sustain itself without energy, and cannot fully perform its functions without a stable supply. The fact is transformers, generators, batteries, and other off-line supplies incur substantial voltage and current variations across time and over a wide range of operating conditions. They are normally noisy and jittery not only because of their inherent nature but also because high-power switching circuits like central-processing units (CPUs) and digital signal-processing (DSP) circuits usually load it. These rapidly changing loads cause transient excursions in the supposedly noise-free supply, the end results of which are undesired voltage droops and frequency spurs where only a dc component should exist. The role of the voltage regulator is to convert these unpredictable and noisy supplies to stable, constant, accurate, and load-independent voltages, attenuating these ill-fated fluctuations to lower and more acceptable levels.

The regulation function is especially important in high-performance applications where systems are increasingly more integrated and complex. A system-on-chip (SoC) solution, for instance, incorporates numerous functions, many of which switch simultaneously with the clock, demanding both high-power and fast-response times in short consecutive bursts. Not responding quickly to one of these load-current transitions (i.e., load dumps) forces storage capacitors to supply the full load and subsequently suffer considerable transient fluctuations in the supply. The bandwidth performance of the regulator, that is, its ability to respond quickly, determines the magnitude and extent of these transient variations.

Regulators also protect and filter integrated circuits (ICs) from exposure to voltages exceeding junction-breakdown levels. The requirement

is more stringent and acute in emergent state-of-the-art technologies whose susceptibility to breakdown voltages can be less than 2 V. The growing demand for space-efficient, single-chip solutions, which include SoC, system-in-package (SiP), and system-on-package (SoP) implementations, drives process technologies to finer photolithographic and metal-pitch dimensions. Unfortunately, the maximum voltage an IC can sustain before the onset of a breakdown failure declines with decreasing dimensions and pitch because as the component density increases, isolation barriers deteriorate.

References, like regulators, generate and regulate accurate and stable output voltages that are impervious to variations in the input supply, loading environment, and various operating conditions. Unlike regulators, however, references do not supply substantial dc currents. Although a good reference may shunt positive and negative noise currents, its total load-current reach is still relatively low. In practice, references supply up to 1 mA and regulators from 5 mA to several amps.

1.2 Linear versus Switching Regulators

A voltage regulator is normally a buffered reference: a bias voltage cascaded with a noninverting op-amp capable of driving large load currents in shunt-feedback configuration. Bearing in mind the broad range of load currents possible, regulators are, on a basic level, generally classified as *linear* or *switching*. Linear regulators, also called *series* regulators, linearly modulate the conductance of a series pass switch connected between an input dc supply and the regulated output to ensure the output voltage is a predetermined ratio of its bias reference voltage, as illustrated in Fig. 1.1*a*. The term "series" refers to the pass element (or switch device) that is in series with the unregulated supply and the load. Since the current flow and its control are

NOTE ON TEXT: To complement and augment the verbal explanations presented in this book, an effort has been made to conform variable names to standard small-signal and steady-state naming conventions. Signals embodying both small-signal and dc components use a smaller-case name with uppercase subscripts, like for instance output voltage v_{OUT}. When referring only to the dc component, all capitals are used, as in V_{OUT}, and similarly, when only referring to small-signal values, the entire name, including subscripts, is in lower case, as in v_{out}. As also illustrated by the previous example, the variables adopt functionally intuitive names. The first letter usually describes the signal type and its dimensional units such as v for voltage, i for current, A or G for amplifying gain, p for power, and so on. The subscript tends to describe the function or node to which the variable is attached, such as "out" for the output of the regulator, "reg" for a regulated parameter, and so on.

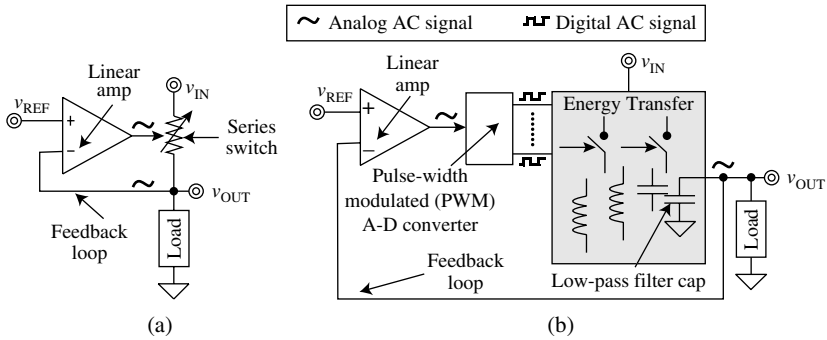


FIGURE 1.1 (*a*) Basic linear and (*b*) switching regulator circuits.

continuous in time, the circuit is linear and analog in nature, and because it can only supply power through a linearly controlled series switch, its output voltage cannot exceed its unregulated input supply (i.e., $V_{OUT} < V_{IN}$).

A switching regulator is the counterpart to the linear solution, and because of its switching nature, it can accommodate both alternating-current (ac) and direct-current (dc) input and output voltages, which is why it can support ac-ac, ac-dc, dc-ac, and dc-dc converter functions. Within the context of ICs, however, dc-dc converters predominate because the ICs derive power from available dc batteries and off-line ac-dc converters, and most loading applications in the IC and outside of it demand dc supplies to operate. Nevertheless, given its ac-dc converting capabilities, switching regulators are also termed *switching converters*, even if only dc-dc functions are performed.

From a circuit perspective, the driving difference between linear and switching regulators is that the latter is mixed-mode with both analog and digital components in the feedback loop (Fig. 1.1*b*). The basic idea in the switching converter is to alternately energize inductors and/or capacitors from the supply and de-energize them into the load, transferring energy via quasi-lossless energy-storage devices. To control the network, the circuit feeds back and converts an analog error signal into a pulse-width-modulated (PWM) digital-pulse train whose on-off states determine the connectivity of the aforementioned switching network. From a signal-processing perspective, the function of the switching network is to low-pass-filter the supply-level swings of the digital train down to a millivolt analog signal, the average of which is the regulated output.

The blocks that normally comprise a dc-dc converter include a PWM controller, which is the combination of an analog linear amplifier and a pulse-width-modulated analog-digital converter, as shown in Fig. 1.1*b*, synchronous and/or asynchronous switches (i.e., transistors and/or diodes), capacitors, and, in many cases, inductors. Many switched-capacitor implementations do not require power inductors,

sometimes making total chip integration possible. These integrated, inductorless converters, however, cannot typically supply the high-current levels the discrete power inductors can, which is why they normally satisfy a relatively smaller market niche in low-power applications.

Switching regulators, unlike their linear counterparts, are capable of generating a wide range of output voltages, including values below and above the input supply. *Buck* converters, for instance, generate output voltages lower than the input supply (i.e., $V_{OUT} < V_{IN}$) while *boost* converters deliver the opposite (i.e., $V_{OUT} > V_{IN}$)—*charge pump* is the name normally applied to an inductorless buck or boost converter. Buck-boost converters, as the name implies, are a combination of both buck and boost circuits and they are consequently capable of regulating output voltages both above and below the input supply. In spite of the apparent flexibility and advantages of switching supplies, however, linear regulators remain popular in consumer and high-performance electronics, as the next subsection will illustrate.

1.2.1 Speed Tradeoffs

Linear regulators tend to be simpler and faster than switching converters. As Fig. 1.1 illustrates, there are fewer components in a linear regulator, which imply two things: simplicity and less delay through the feedback loop, in other words, higher bandwidth and therefore faster response. The PWM controller, and more specifically, the pulse-width-modulated analog-to-digital converter, is generally a relatively laborious block to design, often requiring a clock, comparators, non-overlapping digital drivers, and a saw-tooth triangular-wave generator. For a stable switching converter in negative feedback, the switching frequency is often a decade above the bandwidth of the loop, further limiting its response time to orders of magnitude below the transitional frequency (f_t) of the transistors available in a given process technology. Because of this, and the fact they are relatively complex circuits (i.e., more delays across the loop), dc-dc converters require more time to respond than linear regulators, 2–8 μs versus 0.25–1 μs. The switching frequencies of these devices are between 20 kHz and 10 MHz. Although higher switching frequencies can reduce the ripple content of the output voltage and/or relax the LC-filter requirements, they are often prohibitive because they increase the switching power losses of the converter beyond acceptable limits—increasing power losses demands more energy from the battery and therefore reduces its runtime.

1.2.2 Noise

Switching regulators are noisier than their linear counterparts are, and Fig. 1.1 illustrates this by the presence of digital signals in the ac-feedback path of the circuit. Power switches, which are large devices conducting

high currents, must switch at relatively high frequencies, forcing their driving signals to be fast and abrupt and injecting noise into the energy-storage devices used to supply the loading circuits. The noise is especially prevalent in boost configurations where radio-frequency (RF) noise tends to be worse. Start-stop clock operation, like on-off sleep-mode transitions, further aggravates noise content with low- and high-frequency harmonics.

1.2.3 Efficiency

Switching regulators have one redeeming quality when compared to linear regulators: they are power efficient. The fact is the voltage (and therefore power) across the power switches in a dc-dc converter are far lower (e.g., 10–100 mV) than the voltage across the series pass device of a linear regulator, which is the difference between the unregulated input and regulated output (e.g., 0.3–2 V). Allowing the regulator to dissipate (not deliver) a larger proportion of power results in decreased *power efficiency*, which is an important metric in power-conditioning circuits defined as the ratio of output power P_{OUT} to input power P_{IN}, where the latter comprises delivered output power P_{OUT} and consumed power losses P_{REG}:

$$\eta = \frac{P_{OUT}}{P_{IN}} = \frac{P_{OUT}}{P_{REG} + P_{OUT}} \tag{1.1}$$

Switching regulators commonly achieve efficiencies between 80 and 95%. In the case of linear regulators, quiescent-current flow I_Q and the voltage difference between the unregulated supply V_{IN} and regulated output V_{OUT} limit power-efficiency performance to considerably lower levels,

$$\eta_{Lin-Reg} = \frac{I_{LOAD}V_{OUT}}{(I_{LOAD} + I_Q)V_{IN}} < \frac{V_{OUT}}{V_{IN}} \tag{1.2}$$

where I_{LOAD} is the load current and quiescent current I_Q flows to ground, not the load. The maximum possible efficiency a series regulator can attain is therefore the ratio of the output and input supply voltages, even if its quiescent current nears zero. For instance, the maximum power efficiency a 2.5 V linear regulator can ever achieve while powered from a 5 V input supply is only 50%.

Power efficiency in a linear regulator increases with lower input-output voltage differentials. For instance, if in the above-stated example the regulator drew power from a 3.3 V input supply, the efficiency would have been 76%, or from a 2.8 V input supply, 89%. This characteristic holds true only if average load current I_{LOAD} is considerably

greater than average quiescent current I_Q, which is typical when a full load is presented, but not when the system is idling or asleep. Consequently, when the voltage drop between the unregulated supply and the output is relatively low (i.e., $V_{IN} - V_{OUT} < 0.3$ V), linear regulators are often preferred over their switching counterparts because efficiencies are on par, and the circuit is simpler, less expensive, less noisy, and faster. Their only, though significant, drawback is power efficiency, and if that is not an issue or if equivalent to a switching converter, a linear regulator is best.

Increasing load currents to the point where heat sinks are required is costly. A heat sink increases overhead by requiring an additional on-board component and demanding more real estate on the printed-circuit-board (PCB). A common technique used to circumvent this drawback is to utilize several linear regulators throughout the PCB to split the load and minimize the power dissipated by each regulating IC, or by replacing them with a switching regulator, if performance specification requirements allow, in other words, if more noise in the regulated output is permissible. Another detrimental side effect of high temperature is higher metal-oxide-semiconductor (MOS) switch-on resistances, the results of which are higher conduction losses and consequently lower efficiency performance. In all, as Table 1.1 summarizes, linear regulators are simpler and faster and produce lower noise levels, but their relatively limited efficiency performance, however, constrains them to lower power applications and dedicated supply systems. Switching regulators, on the other hand, may enjoy more efficiency but the loads they sustain must tolerate higher noise levels, which is why power to high-performance analog subsystems is normally channeled by way of linear regulators.

Linear Regulators	Switching Regulators
Output range is limited ($V_{OUT} < V_{IN}$)	√ Output range is flexible ($V_{OUT} \leq V_{IN}$ or $V_{OUT} \geq V_{IN}$)
√ Simple circuit	Complex circuit
√ Low noise content	High noise content
√ Fast response	Slow response
Limited power efficiency ($\eta < V_{OUT}/V_{IN}$)	√ High power efficiency ($\eta \approx$ 80–95%)
Good for low-power applications	Good for high-power applications

Check mark "√" denotes positive attributes (i.e., advantages).

Table 1.1 Comparing Linear Regulators to Their Switching Counterparts

1.3 Market Demand

1.3.1 System

Both linear and switching regulators claim their place in today's market. Systems like desktop and laptop microprocessors not only demand substantial amounts of clock-synchronized currents but also low supply voltages. Such systems reap the benefits of power efficient dc-dc converters. Circuit blocks serving purely analog functions, on the other hand, cannot sustain the noisy supplies generated by switching regulators and therefore exploit the low-noise and cost-effective advantages of linear regulators. These analog circuits are inherently more sensitive to noise originated in the supply rails than digital blocks, which is why they require "cleaner" power supplies.

Today's growing market demand for portable electronics, like cellular phones, portable digital assistants (PDAs), MP3 players, laptops, and the like requires the use and coexistence of both linear and switching regulators, since both accuracy and power efficiency are paramount. In these applications, the integrated power management circuit drives noise-sensitive circuits from noisy and variable input-supply voltages. A dc-dc converter, under these conditions, steps down the input supply to a lower voltage level, generating in the process a regulated but noisy supply voltage (e.g., v_{NOISY}), as shown in Fig. 1.2. A linear regulator then draws power from this noisy supply to generate a low-noise, ripple-rejected output (e.g., v_{CLEAN}), which is now compatible with high-performance, noise-sensitive ICs. The voltage across the linear regulator is therefore low enough to limit its power losses (e.g., 2.4–1.8 V in the figure). The purpose of the switching regulator is therefore to down-convert as much of the input voltage as possible to save power, since it is more power efficient than the linear regulator. The function of the linear regulator is to filter the noise and generate the noise-free supply that the system demands.

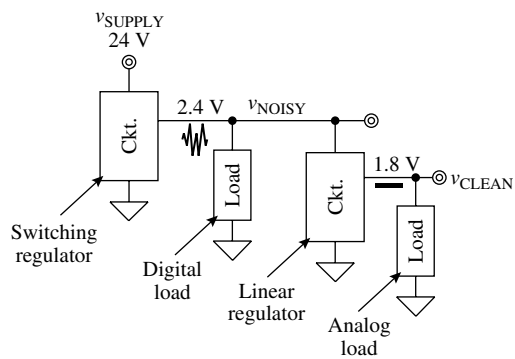


FIGURE 1.2
Sample low-noise power management system.

Similar operating conditions to those just described arise in many other mixed-signal applications, where active power supply decoupling is necessary to reduce and suppress noise. Systems demanding high-output voltages from low input-supply voltages, as is the case for single- or dual-battery packs (e.g., 0.9–1.5 V), require the use of boosting dc-dc converters. As in Fig. 1.2, series regulators may still be required to suppress the switching noise generated by the switching regulator.

1.3.2 Integration

The mobile market's impact on the demands of regulators is pronounced. Because of high variations in battery voltage, virtually all battery-operated applications require regulators. What is more, most designs find it necessary to include regulators and other power-supply circuits in situ, on-chip with the system to save printed-circuit-board (PCB) real estate and improve performance. This trend is especially prevalent in products that strive to achieve or approach the fundamental limits of integration in the form of *system-on-chip* (SoC), *system-in-package* (SiP), and *system-on-package* (SoP) solutions. Limited energy and power densities are the by-products of such a market, requiring circuits to yield high power efficiency and demand low quiescent-current flow to achieve reasonable lifetime performance.

1.3.3 Operational Life

Current efficiency (η_I), which refers to a proportionately lower quiescent current when compared against the load current, is also vital. In particular, quiescent current (I_Q) must be as low as possible during zero-to-low loading conditions because it amounts to a significant portion of the total drain current of the battery. During heavy loading events, on the other hand, higher quiescent current is acceptable because its impact on total drain current and therefore battery life is miniscule, which is why current efficiency, and not absolute quiescent current, is important:

$$\eta_I = \frac{I_{LOAD}}{I_{TOTAL}} = \frac{I_{LOAD}}{I_{LOAD} + I_Q} \tag{1.3}$$

Ultimately, the load alone determines the lifetime of the battery during high load-current conditions and the regulator's quiescent current during zero-to-low loading events.

The capacity of a battery, defined in amp-hours, and the average drain current sets the battery life of the electronic system:

$$\text{Life[h]} = \frac{\text{Capacity[Ah]}}{I_{DRAIN(ave)}} = \frac{\text{Capacity[Ah]}}{I_{Q(ave)} + I_{LOAD(ave)}} \tag{1.4}$$

This relationship, coupled with the fact that the majority of portable devices idle most of the time, implies that battery life is a strong function

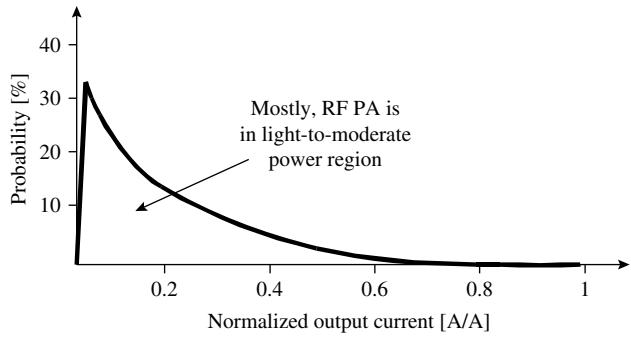


FIGURE 1.3 Probability-density function (PDF) of a typical radio-frequency (RF) power amplifier (PA) in a portable CDMA handset application.

of low load-current conditions, that is of $I_{Q(ave)}$. A cellular phone, for instance, mostly idles (i.e., *alert* but not in *talk* mode) and consequently requires only a fraction of its peak power most of the time, as depicted by the probability-density function (PDF) in Fig. 1.3. As shown, the region of highest probability is the zero-to-moderate load-current range, which is where drain current is mostly comprised of quiescent-current flow I_Q:

$$I_{DRAIN(ave)} = \int (I_{LOAD} + I_Q) \cdot PDF \cdot dI_{LOAD} \approx I_Q \tag{1.5}$$

1.3.4 Supply Headroom

Battery power and state-of-the-art process technologies also imply low-voltage operation. Today's most popular secondary (i.e., rechargeable) battery technologies are lithium ion (Li Ion), nickel cadmium (NiCd), and nickel metal hydride (NiMH), the first of which ranges from 2.7 V when completely discharged to 4.2 V when fully charged and the latter two from 0.9 to 1.7 V. Microscale fuel cells have even lower voltages, approximately at 0.4–0.7 V per cell. Ultimately, the variable nature of these relatively low-voltage technologies superimposes stringent requirements on the regulator, limiting their supply headroom voltage and their available dynamic range to considerably low levels.

Low-voltage operation is also a consequence of advances in process technologies. The push for higher packing densities forces technologies to improve their photolithographic resolution, fabricating nanometer-scale junctions, which have inherently lower junction breakdown voltages. A typical 0.18 μm CMOS technology, for instance, cannot sustain more than roughly 1.8 V. Additionally, since financial considerations limit the complexity of the process, that is, the number of masks used to fabricate the chip and therefore the variety of devices available, only *vanilla* (standard) CMOS and stripped

BiCMOS process technologies are most desirous, which translates to less flexibility for the designer.

A low-voltage environment is restrictive for an analog IC designer. Many traditional design techniques are prohibitive under these conditions, limiting flexibility and sometimes system performance. Cascoding devices, emitter and source followers, and Darlington-configured bipolar-junction transistors costs, for instance, which are useful for increasing gain, bandwidth, and current-driving capabilities, require additional voltage headroom, which is a precious commodity in battery-operated devices. Low voltages also imply increased precision, pushing performance down to fundamental limits. A 1% 1.8 V regulator, for example, must have a total variation of less than 18 mV, which includes 5–12 mV of load and line regulation effects, extended commercial temperature extremes (e.g., –40°–125°C), process variations, noise and variations in the supply, and so on. What is more, since dynamic range suffers in a low-voltage setting, the demand for improved percent accuracy increases to sub-1% levels. These issues typically give rise to more complex and usually more expensive ICs (i.e., more silicon real estate and/or more exotic circuit and/or process technologies), encouraging the designer to be more resourceful and innovative.

1.4 Batteries

It is important to comprehend fully the environment in which many, if not most, linear regulators find a home. Typical parameters to consider in a battery, from an IC designer's perspective, are capacity, cycle life, internal resistance, self-discharge, and physical size and weight. Cycle life refers to the number of discharge-recharge cycles a battery will endure before significantly degrading its capacity, that is, its ability to store energy. There are several types of batteries, ranging from reusable alkaline and nickel cadmium (NiCd) to lithium ion (Li Ion) and lithium-ion polymers. Unfortunately, however, in spite of advances in battery technology, there is no one-battery solution for *all* possible applications.

1.4.1 Early Batteries

Most portable electronics today use either nickel- or lithium-based chemistries to power their systems. Reusable alkaline and lead-acid batteries are not appropriate for high-performance applications because of cycle-life and integration limits. Alkaline batteries, for instance, have long shelf lives but suffer from short cycle lives and low-power densities. They are therefore best suited for a range of consumer electronics and gadgets that demand reliable operation on an infrequent basis, like flashlights. Lead-acid batteries are economical and good for high-power applications, but they are bulky, which

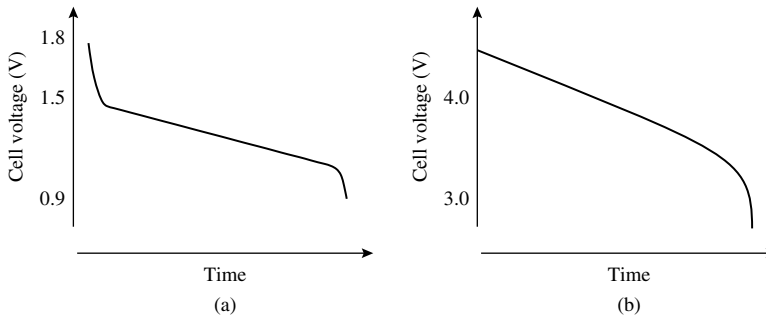


FIGURE 1.4 (*a*) NiCd and NiMH and (*b*) Li Ion discharge curves under a constant load current.

is why only macroscale applications like the automobile industry benefit most from their features, unlike the portable, handheld electronics industry.

Early cellular phones used nickel-based batteries: nickel cadmium (NiCd) and nickel-metal hydride (NiMH). These nickel-based solutions suffer from a phenomenon known as *cyclic memory*. To prevent the negative effects of cyclic memory, which amounts to crystalline formation and consequently higher self-discharge rates, periodical discharge-charge cycles are necessary. Figure 1.4*a* illustrates the typical discharge curve of these nickel-based devices, showing how most of the usable energy is in the 1–1.5 V range.

NiCd batteries, which contain toxic metals, are the predecessors of the NiMH solution, which is more environmentally friendly and yields slightly higher energy densities and lower memory effects. The advantages of NiMH, however, come at the cost of other performance metrics. Figure 1.5, for instance, illustrates how NiCd eventually outperforms NiMH in almost every way over its cycle life. Not only is internal

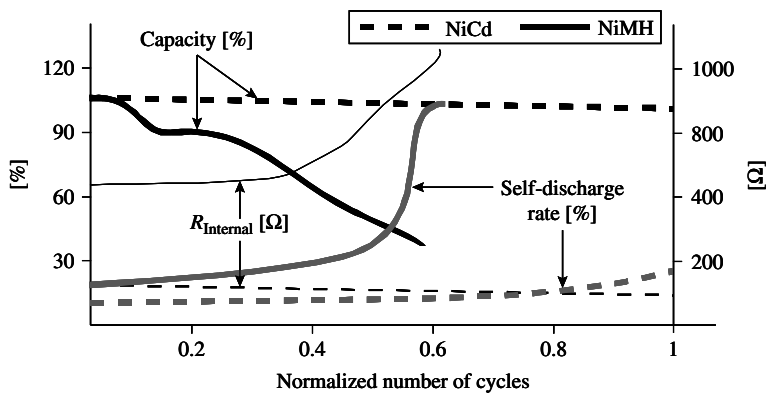


FIGURE 1.5 Comparative performance of NiCd and NiMH batteries.

resistance lower for NiCd batteries but capacity, internal resistance, and self-discharge remain relatively constant throughout their full cycle life, roughly 1500 cycles' worth. NiMH batteries perform well at first but quickly degrade within 20% of the NiCd's full life after a limited number of charge cycles, giving them roughly half the usable life of a NiCd battery. NiMH devices, nevertheless, continue to appeal to the electronics industry because it is environmentally friendly, and perhaps more importantly, from a marketing perspective, because the life expectancy of most microelectronic products today is relatively short, on the order of a year, limiting the number of recharge cycles to within the range of NiMH technologies. Increased volume sales will therefore spark innovation and advancements in technology and consequently reduce the cost of NiMH batteries to more competitive levels with respect to NiCd technologies.

1.4.2 Li Ion Batteries

Next in the evolutionary chain of rechargeable energy-storage devices are the lithium-ion (Li-Ion) batteries, which have the highest energy density levels, when compared to the nickel-based chemistries, and they do not contain toxic metals or the dreaded memory effects. They exhibit relatively constant capacity and internal-resistance performance over most of their entire cycle life, which extends up to approximately 1000 cycles. Their self-discharge rates are miniscule when compared against nickel-based chemistries. All these advantages come at the cost of technology, that is, at a higher dollar premium (roughly twice the cost), which is not to say the price will not decrease in the near future because it will, as more and more of them are sold, benefiting from economies of scale. Because of these reasons, most cellular phones, laptops, and other portable consumer electronics use Li Ions.

At a slightly higher cost, Li-Ion polymers offer similar performance, but with the ability to conform into thinner and smaller packages, which is especially useful in handheld, wearable, and wireless-sensor applications. Ultimately, most Li-Ion technologies conform to the discharge curves shown in Fig. 1.4*b*, where most of its useful energy falls within the 2.7–4.2 V region. Because of the sensitive nature of the chemistry, charging or discharging them beyond maximum- and minimum-rated limits (4.2–2.7 V, normally) causes irreversible and sometimes catastrophic effects, which is why the charging circuit for these batteries is often more complex than for other technologies.

1.4.3 Fuel Cells

Li Ions are mainstream, but not ideal. They do not store enough energy per unit weight or space for moderate-power, microscale systems, which is why fuel cells (FCs), energy harvesters, and nuclear batteries are the subject of attention in the world of research. Figure 1.6 illustrates a Ragone plot of various energy-storage devices, depicting

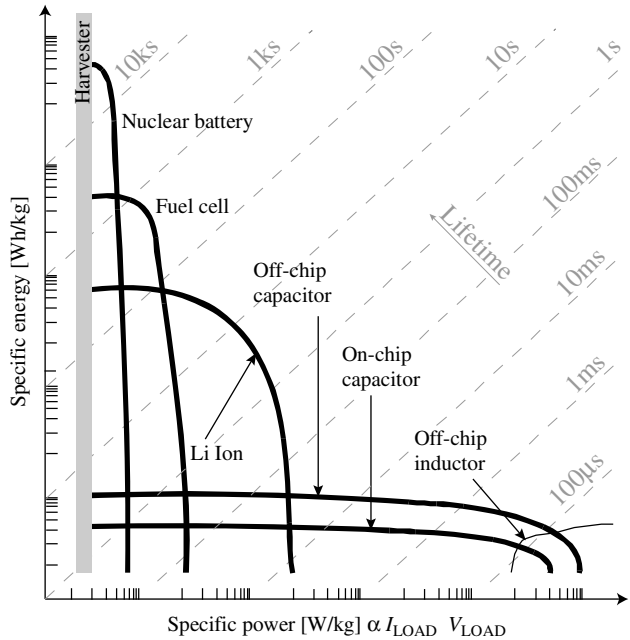


Figure 1.6 Ragone plot: comparative energy-power performance of various energy-storage technologies.

their respective energy-to-power relationships. FCs, for instance, like nuclear batteries and microscale energy harvesters, have higher energy at low-power levels, whereas Li-Ion technology has higher energy at higher power levels; in other words, given similar volume constraints, a Li Ion outlives an FC under high-power conditions, and vice versa. What is more, FCs are inherently slower to respond to load changes than Li-Ion batteries, and with inherently lower voltages (0.4–0.7 V). Ultimately, none of these technologies are ideal. Researchers are therefore looking to improve each of these technologies separately and, at the same time, leveraging their complementary features in compact hybrid solutions.

1.4.4 Nuclear Batteries

In the battle for maximum energy storage, nuclear batteries overwhelm the others at lower power levels, except for energy harvesters, which provide essentially boundless energy. The main drawback is the radioactive nature of the material used to build it, and its implied safety and containment requirements. The secondary but equally relevant disadvantage is its low power. The heat this technology generates, however, can provide long-term fuel for thermoelectric generators—for a decade, for instance. Similarly, the electrons emitted by decaying isotopes can be used to establish electric fields across parallel and

mechanically compliant piezoelectric plates whose attractive force bends the material and therefore generates current flow on contact. The emitted electrons can also be used to generate electron-hole pairs in pn-junction devices, much like photons are used in photovoltaic solar cells; these devices are called β-*voltaic batteries* and enjoy the same chip-integration benefits of solar cells. In the end, however, safety and the cost of these isotopes prevent the penetration of these batteries into the marketplace, but research certainly continues, as their energy content is unequally high.

1.4.5 Harvesters

Last, but certainly not least, in the race for long battery life are energy harvesters. These electric generators extract fuel energy from the environment, from motion, heat, and pressure, and so on. These generators are at the forefront of research and solutions have yet to mature, but they promise to compete with nuclear batteries in the race for extended, perhaps even perpetual, life. Microscale harvesters fabricated with microelectromechanical systems (MEMS) technologies are compatible with ICs and may therefore see the light of day, but it is still too early to tell.

1.5 Circuit Operation

1.5.1 Categories

Power

There are various types of linear regulators catering to an assortment of different applications. Generally, the most obvious distinction between them is power level. Low-power regulators, for instance, normally supply output currents of less than 1 A, which is typical of many portable and battery-powered electronics, and high-power regulators source higher currents for automotive, industrial, and other applications of the ilk. High-power linear regulators, however, are quickly losing ground to their switching counterparts because of their power-efficiency performance deficit. A system will more than likely use a master switching converter and sprinkle linear regulators at various load points, conforming to a *point-of-load (PoL) regulation* strategy. PoL provides better overall performance in the form of dc and ac (noise content) accuracy at each individual load. At present, linear regulators find most of their market in the sub-300-mA region.

Compensation

Within the context of circuit architecture, two other major classifications exist: *externally* and *internally compensated* structures. A capacitor used to stabilize the negative feedback loop of the regulator (i.e., to

setting the dominant low-frequency pole) that is connected between any two of the input/output (I/O) pins of the IC (i.e., input supply, ground, and regulated output) is said to compensate the circuit externally. When an internal node is used for connecting a compensation capacitor, the circuit is said to be internally compensated. In this latter case, the capacitor hanging off the output cannot exceed a specified maximum value because the output is a parasitic pole and increasing its capacitance pulls the pole it sets closer to in-band frequencies, possibly giving rise to unstable conditions. Output capacitors for externally compensated circuits, on the other hand, set the dominant low-frequency pole at the output and must therefore be sufficiently large (i.e., exceed a specified minimum) to guarantee stable conditions.

Circuit applications normally require an output capacitor to suppress transient load excursions. Consider a linear regulator incurs some finite delay before fully responding to quick load dumps, during which time the output capacitor sources or sinks the difference. A larger-output capacitor therefore droops less and yields better transient-response performance (i.e., less output voltage variation) in the presence of load-dump events demanding substantial currents:

$$\Delta v_{OUT} = \frac{\Delta i_{LOAD} t_{delay}}{C_{OUT}} \alpha \frac{1}{C_{OUT}} \tag{1.6}$$

For instance, in the presence of a 50-ns, 1–11 mA load dump, a regulator with a bandwidth of 100 kHz and a 0.47 μF output capacitor allows its output voltage to droop approximately 10 mA$(3/2\pi \cdot 100$ kHz $\cdot$ 50 ns$)/0.47$ μF or 101 mV before responding and sourcing the full load. Unlike internally compensated regulators, the output capacitor's requirements for an externally compensated circuit align well with transient-noise suppression, which is why users prefer externally compensated circuits, in spite of the dollar and printed-circuit-board (PCB) real-estate costs associated with off-chip capacitors.

The advent of *system-on-chip* (SoC) and *system-in-package* (SiP) integration is slowly changing the trend from externally to internally compensated schemes, however. As the IC absorbs more circuits and regulators into its common silicon substrate, external capacitors are increasingly more difficult to accommodate. Externally compensated circuits (where the dominant low-frequency pole is at the output) are therefore appearing under the guise of internally compensated ICs simply because the output capacitor, although still hanging off the output node, is in the IC and no off-chip capacitor is required. Although advertised as internally compensated structures, for the sake of the discussions in this text, given their technical implications on the circuit and feedback loop, linear regulators whose dominant low-frequency pole is at the output are said to be externally compensated, whether the compensation capacitor is on-chip or not.

Dropout

Linear regulators are also classified as low or high dropout (LDO or HDO), which refers to the minimum voltage dropped across the circuit, in other words, the minimum difference between the unregulated input supply and the regulated output voltage (V_{DO} in Fig. 1.7). This voltage is important because it represents the minimum power dissipated by the regulator, since the power lost is dependent on the product of the load current and this dropout voltage. *Low-dropout* (LDO) *regulators* consequently dissipate less power than their higher dropout counterparts and have therefore enjoyed increasing popularity in the marketplace, especially in battery-operated environments. Linear regulators with dropout voltages below 600 mV belong to the low-dropout class, but typical dropout voltages are between 200 and 300 mV.

Figure 1.7 also illustrates the three regions of operation of a linear regulator: *linear*, *dropout*, and *off* regions. When the circuit is operating properly, that is to say, when it regulates the output with some finite and nonzero loop gain, the regulator is in the linear region. As input voltage v_{IN} decreases, past a certain point, one of the transistors in the loop enters the triode region (or low-gain mode) during which time the circuit still regulates the output, albeit at a lower loop gain and consequently with some gain error. As v_{IN} decreases further, the loop gain continues to fall until it becomes, for all practical purposes, zero, when it reaches its driving limit. At this point, the regulator enters the dropout region and the power switch, given its limited drive, operates like a switch because it supplies all the current it can to maintain the highest possible output voltage. The voltage difference between v_{IN} and v_{OUT} in this region is dropout voltage V_{DO}, and although V_{DO} is at first approximately constant, as though it were a resistive ohmic voltage drop, it tends to increase with decreasing values of v_{IN}—datasheets often quote the equivalent resistance during the mostly linear

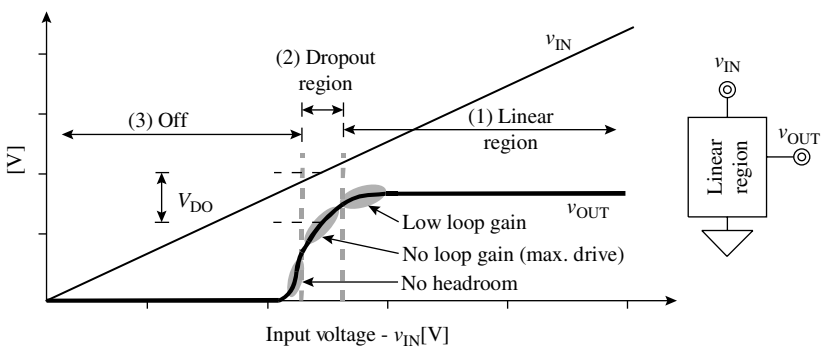


Figure 1.7 Typical input-output voltage characteristics of a linear regulator.

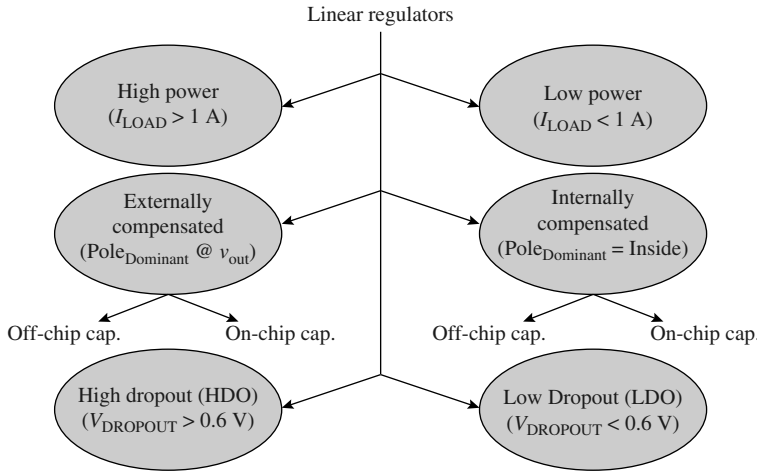


Figure 1.8 Linear regulators and their corresponding subclasses.

portion of the dropout region. The off region is where the circuit reaches its headroom limit, when the input supply voltage is too low for the transistors to work properly, and more specifically, for the negative feedback control loop to process that v_O is below its target and keep some drive applied to the power switch.

In summary, linear regulators can be high or low power, externally or internally compensated, and high or low dropout, as depicted in Fig. 1.8. Most linear regulators of interest nowadays fall in the low-power regime. Because of improved efficiency performance, switching converters now fill a wider high-power market segment, from which several low-power subsystems derive power directly or via series linear regulators. Internally compensated linear regulators with only on-chip capacitors, though not as popular today, are highly desirable because of their low PCB real-estate and dollar-cost implications. They respond poorly to quick load-dump events, however, and therefore suffer from degraded ac accuracy performance. Their popularity is growing in lower power application-specific ICs (ASICs), however, where one or several linear regulators share the silicon die with their respective loading components.

Convolving the general characteristics just described with the increasing market demand for portable, battery-powered electronics birth a niche market for low power, internally compensated, (LDO) regulators. This consumer market segment demands the small-footprint and extended-life solutions internally compensated circuits with low-dropout voltages enable. Decreasing the voltage dropped across the regulator (i.e., the difference between the unregulated supply and the regulated output) reduces the power dissipated by the

regulator (P_{LDO}) and therefore increases efficiency, which ultimately translates to extended single-charge operational life:

$$P_{LDO} = I_{QUIESCENT}V_{IN} + I_{LOAD}V_{DO} \geq I_{LOAD}V_{DO} \tag{1.7}$$

Dropout voltage and load current, as seen in the relation above, define the minimum power a linear regulator can *ever* dissipate under moderate-to-full loading conditions. The dropout voltage specification is therefore of paramount importance.

From the viewpoint of battery-supplied systems, LDO regulators relax the headroom limits and increase the dynamic range and signal-to-noise ratio (SNR) performance of their loading circuits, which are normally the first to degrade when the input supply voltage decreases. For instance, if an almost drained Li Ion (i.e., V_{IN} is roughly 2.7 V) powers an LDO regulator with a dropout of 0.2 V, which then powers an operational amplifier whose noise floor is 10 mV, the operational amplifier would have to operate under a 2.5 V regulated supply, which is, for all practical purposes, already a low supply voltage. All processing must therefore fall within a 2.5 V window, which qualitatively sets the maximum possible dynamic range of the circuit:

$$\text{SNR} = \frac{\text{dynamic range}}{\text{noise floor}} = \frac{2.5\text{ V}}{10\text{ mV}} = 250 \tag{1.8}$$

Had the regulator been a high-dropout (HDO) device with a dropout of 0.7 V, the operational amplifier would have had to work under a 2.0 V regulated supply, which imposes serious restrictions on the amplifier circuit, not to mention its ability to process analog information, now within a 2.0 V window and a signal-to-noise ratio of less than 200. Additionally, the output voltage, as depicted in Fig. 1.7, exhibits less variation over the full span of the unregulated supply range with an LDO circuit than with an HDO regulator, since the onset of dropout is lower for LDO regulators.

Low-dropout voltages are necessary in many applications, from automotive and industrial to medical. The automotive industry, for instance, exploits the low-dropout characteristics of LDO regulators during cold-crank conditions, when the car-battery voltage is between 5.5 and 6 V and the regulated output must be around 5 V, requiring a loaded dropout voltage of less than 0.5 V. The increasing demand is especially apparent, however, in mobile battery-operated products such as cellular phones, pagers, camera recorders, and laptops. Such space-efficient designs only use a few battery cells, thereby necessarily decreasing the available input supply voltage. In these cases, LDO regulators are capable of supplying relatively high output voltages. For instance, if two NiCd cells, which have an approximate range of 1.8–3 V, power an HDO regulator, the maximum output voltage over the life span of the battery is less than 1.2 V. This output voltage

is not sufficiently high to meet the headroom requirements of most analog circuits. What is more, the restriction is prohibitive for single battery-cell conditions.

1.5.2 Block-Level Composition

A regulator is mainly comprised of a control loop whose function is to monitor and control its output to remain within a small window of a target reference value, irrespective of its environment and its operating conditions. A regulator circuit must therefore sense the output, compare it against a reference, and use the difference to modulate the conductance between the input supply and the regulated output. In the case of a voltage regulator, a feedback network senses the output, as shown in Fig. 1.9, and feeds it to an error amplifier, whose function is to compare it against a reference voltage and generate an error signal that modulates the conductance of a pass device. The circuit is essentially a noninverting operational amplifier with a dc reference voltage at its noninverting input.

There are two major blocks in a voltage regulator: a voltage reference and the control loop. The latter is comprised of (1) an error amplifier to sense and generate a correcting signal, (2) a feedback network to sense the output, and (3) a pass device to mediate and conduct whatever load current is required from the unregulated input supply to the regulated output. The control loop, in essence, reacts to offset and cancel the effects of load current, input voltage, temperature, and an array of other variations on the output. The reference

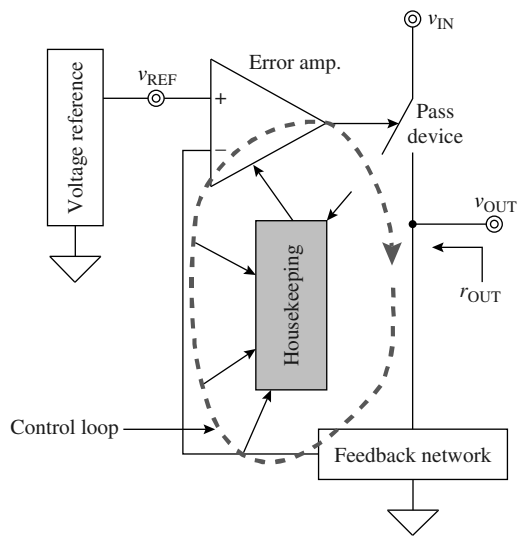


Figure 1.9 General block-level composition of a linear regulator.

block provides a stable dc-bias voltage that is impervious to noise, temperature, and input-supply-voltage variations. The reference's current-driving capabilities are for the most part severely limited, as mentioned earlier in the chapter.

Housekeeping functions are essential to the overall health of the device. They ensure the system operates safely and reliably, protecting the regulator from extreme adverse conditions. Plausible destructive scenarios include exposure to overcurrent, overvoltage, overtemperature, short-circuit, and electrostatic discharge (ESD) events. Housekeeping circuits can also have application-specific features like enable-disable and soft- or slow-start functions for power-moding a system.

1.5.3 Load Environment

Composition

Figure 1.10 illustrates the typical operating environment of a linear regulator IC. The effective loading elements include the parasitic devices associated with the package of the IC, filter capacitors, PCB, and loading circuits. The model presented assumes that the effective load of the regulator begins at the pin and not at the bond-pad because that is the regulation point, where the sense-feedback node is connected. If there is no bond wire allocated to the sense node, the effective load starts at the bond pad, where both the power device and sense node converge. In the case the regulation point is at the pin, as in Fig. 1.10, the parasitic bond wire inductance (L_{BW}) and resistance (R_{BW}) are part of the regulator, and the feedback loop consequently combats to mitigate their adverse effects on the output; in other words, the loop regulates the output against any variations in L_{BW} and R_{BW}.

Output capacitor C_{OUT} suppresses the transiently induced voltage variations on the output, as mentioned earlier, and for the case of externally compensated regulators, sets the dominant low-frequency pole of the controlling negative-feedback loop. The output capacitance is normally on the order of several microfarads, as is for input

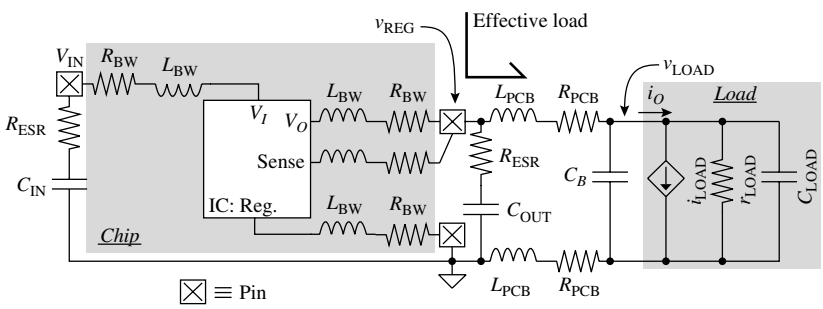


FIGURE 1.10 Typical loading environment of a linear regulator IC.

filter capacitor C_{IN}, which is also used to suppress transient noise in light of the parasitic effects prevalent in practical batteries, such as finite output resistances and limited response times. Because of lower cost, tantalum capacitors are normally used. These devices, as all others to various degrees, exhibit a parasitic equivalent-series resistance (ESR) denoted in the figure as R_{ESR}, the value of which can be up to several ohms.

Ceramic and ceramic multilayer chip (CMC) capacitors have low ESR values and consequently lend themselves for higher frequency applications and improved transient-noise suppression, which is why designers often place them near the load, as bypass capacitors (C_B in Fig. 1.10). These higher cost devices are normally in the nanofarad region and, in a multiload environment, sum to less than 1 µF. The inspiration behind placing them at the load is point-of-load (PoL) performance such that they supply most of the almost instantaneous load current and ease the drooping effects of the same, supplying the necessary current at the load. Without these high-frequency capacitors, tantalum output capacitor C_{OUT} would have delivered the current through the series parasitic resistors and inductors introduced by C_{out} and the PCB (R_{PCB} and L_{PCB}).

Capacitors also exhibit an equivalent-series inductance (ESL), the magnitude of which is typically less than 5 nH. The effects of this parasitic inductance are often negligible in low-power and low-bandwidth applications. A 10 mA load dump, for example, through a 5 nH inductor in 1 µs produces an ESL voltage of 50 µV ($L\Delta i/\Delta t$ = 5 nH · 10 mA/1 µs). On the other hand, a 100 mA load dump that happens in 50 ns across a 5 nH inductor produces an ESL voltage of 10 mV (5 nH · 100 mA/50 ns), which on a 1 V output constitutes a 1% variation, not including the effects of load and line regulation, capacitor transient droop, temperature, and process variations on the output.

In general, all parasitic devices present in the power path of the regulator cause negative effects on dc, transient, and efficiency performance. Series parasitic bond wire and PCB resistors R_{BW} and R_{PCB} introduce load-dependent series dc ohmic voltage drops and power losses, producing a lower-than-anticipated voltage at the load and higher than expected power losses. Parallel multiple bond wires appeal to the designer because they produce lower series resistances and inductances. For similar reasons, loads that are close to the regulator (with short and wide PCB traces) also produce lower voltage drops and power losses.

Point of Load (PoL)

From the perspective of regulation performance, it is best to sense and regulate the load voltage at the point of load (PoL), not at the output of the power-pass device of the regulator, which is why a *star* connection at the load with a dedicated sense pin yields the best results. The objective is to decouple the regulation point from the

parasitic voltage drops in the power-conducting path, allowing the loop to regulate the load against the reference more accurately. Since the sense node neither sinks nor sources current, no series voltage drops exist in its path. This star connection, where current- and non-current-carrying signal paths belonging to the same node converge at one single point and nowhere else, is also called a *Kelvin* connection. Many applications, however, cannot afford to dedicate a pin for this purpose and resign themselves to a separate sense pad but with a common output-sense pin, as shown in Fig. 1.10. If the application also prohibits the use of multiple bond pads, the sense node is connected to the output at the bond pad, and nowhere else inside the IC, via a dedicated sense path. The objective, in general, is to push the sensing and regulation point as close to the load as the application and technology allow.

The Load

The actual "load" is difficult to model because of its unpredictable nature—the designer is often unaware of what will ultimately load the regulator, except in application-specific cases. As it applies to the regulator, however, dc current I_{LOAD}, equivalent load resistance r_{LOAD}, and equivalent load capacitance C_{LOAD} are the most important parameters because they set the biasing condition of the regulator and the small-signal loading impedance of the same, which affects its stability conditions. Load current i_{LOAD} spans the maximum range specified for the regulator (e.g., 1–50 mA) and incurs the worst-case load dumps for the system during transient conditions, which amounts to the maximum possible load step in the shortest time possible (e.g., 1– 25 mA in 100 ns).

Not knowing the exact nature of the load makes it impossible to predict r_{LOAD} accurately, yet its impact on stability and circuit requirements can be profound. If a low-power operational amplifier whose lowest impedance path to ground may be a diode-connected transistor (with small-signal resistance $1/g_m$) in series with an active load (with relatively larger small-signal resistance r_{ds} or r_o) loads the regulator, the equivalent-load resistance would be on the order of tens to hundreds of kilo-ohms. High-power amplifiers, on the other hand, deliver substantial currents to low-impedance outputs, normally subjecting the supply transistor to its triode region (i.e., low-resistance switch mode) and in series with the low-impedance output. The slewing amplifier therefore establishes a sub-kilo-ohm path from input supply to ground. In the case of digital circuits, like inverters and other CMOS gates, both pull-down and push-up transistors simultaneously conduct shoot-through current during transitions. Although these transitions are short, the equivalent resistance from the supply to ground is the average series combination of two switch-on resistors, both of which are considerably low in value. In the end, r_{LOAD} may span a wide range of resistances.

The designer, for reliability concerns, must therefore consider all extreme conditions: (1) load is purely resistive (i_{LOAD} is zero and $r_{LOAD} = V_{OUT}/I_{LOAD}$) and (2) load is only a current sink (r_{LOAD} is infinitely large or altogether removed). For example, the worst-case (extreme) r_{LOAD} and i_{LOAD} combinations of a 1–50 mA 2.5 V LDO are (1) 2.5 kΩ (2.5 V/1 mA) and 0 mA, (2) 50 Ω (2.5 V/50 mA) and 0 mA, (3) infinite resistance and 1 mA, and (4) infinite resistance and 50 mA, respectively. Simply assuming the load is purely resistive may be unrealistically optimistic or pessimistic. In the case of internally compensated LDOs, for instance, whose output pole is parasitic to the system, a purely resistive load places the output pole at optimistically higher frequencies. Subjecting this LDO to a higher impedance load pulls the output pole to lower frequencies, compromising the stability of the system. Similarly, assuming the load is purely active, that is, only a current sink, may be unrealistically optimistic in the case of externally compensated LDOs, where the dominant low-frequency pole is at the output and a high-impedance load optimistically places this pole at lower frequencies. A lower impedance load pushes the output pole to higher frequencies, closer to the parasitic poles of the system, where stability may be compromised.

Equivalent load capacitance C_{LOAD} is often negligible when compared against bypass and output capacitors C_B and C_{OUT}. This is especially true for standard commercial-off-the-shelf (COTS) LDO ICs and moderate power LDOs because their output capacitors are necessarily large. System-on-chip (SoC) applications, however, do not always enjoy this luxury because C_{OUT} is on-chip and therefore considerably smaller than their off-chip counterparts. Some SoC implementations, in fact, rely entirely on C_{LOAD} for stability and transient response, altogether eliminating the need for C_{OUT}—this is equivalent to using C_{LOAD} as C_{OUT}. As with r_{LOAD}, C_{LOAD} in an SoC environment can play a pivotal role in establishing the stability conditions of the system, which is why the designer must consider all possible values. However, in the presence of a substantially larger C_{OUT} and C_B combination, C_{LOAD} is less important.

1.5.4 Steady-State and Transient Response

Simplified Model

Before subjecting *a transient load-dump event* the model presented in Fig. 1.10, it is helpful and convenient, for the sake of design and insight, to simplify the schematic to those components whose effects dominate the dc, frequency, and transient response of the system. For one, load and bypass capacitors C_{LOAD} and C_B are in parallel and can therefore conform to a single equivalent bypass capacitance, as denoted by C_B' in the now simplified model of Fig. 1.11—given the typical larger values of C_B relative to C_{LOAD}, C_B' normally simplifies to C_B. Because the current flowing through the sense path is negligibly small, the

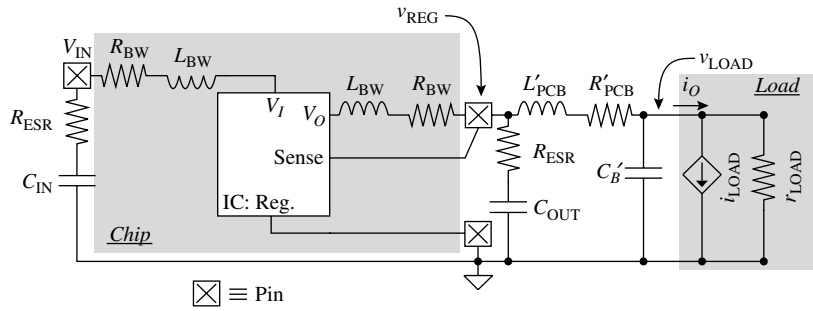


Figure 1.11 Simplified operating environment of a linear regulator IC.

voltage drops associated with its relevant series parasitic bondwire resistance and inductance are practically zero, which is why these components no longer form part of the model. Similarly, when compared with load current i_{LOAD}, LDO's ground current, which is normally just the quiescent current of the regulator, is small; as a result, the adverse effects of i_{LOAD} on the load overwhelm those of this current so its parasitic devices are also removed from the simplified model. Lastly, the series-parasitic elements in the supply and ground-return paths of the load in the PCB have an aggregate effect on load voltage v_{LOAD} relative to regulated output V_{REG}—neglecting the effects of the load's small-signal output resistance r_{LOAD}:

$$v_{LOAD} = v_{REG} - i_{LOAD}(R_{PCB\_SUPPLY} + R_{PCB\_GND}) - \frac{di_{LOAD}}{dt}(L_{PCB\_SUPPLY} + L_{PCB\_GND}) \tag{1.9}$$

and can therefore conform and simplify to a single set of lumped elements,

$$v_{LOAD} = v_{REG} - I_{LOAD}R'_{PCB} - \left(\frac{di_{LOAD}}{dt}\right)L'_{PCB} \tag{1.10}$$

where R'_{PCB} and L'_{PCB} are the lumped total supply and ground-return PCB path series parasitic resistance and inductance of the load, as shown in Fig. 1.11.

DC Variations

Steady-state changes in load current I_{LOAD}, as noted in the previous equation, decrease dc load voltage V_{LOAD} below its targeted regulated value V_{REG}. Had the sense node been connected to the output at the bond pad, the effect would have been increased, adding R_{BW} to the existing R_{PCB}'. Conversely, dedicating a pin to the sense node and connecting it to the load *at the load* (i.e., Pol) would have eliminated the

adverse effects of both R_{BW} and R'_{PCB}. When the circuit enters the dropout region, however, the voltages across these parasitic devices effectively increase the dropout voltage of the LDO, again, degrading its overall performance, irrespective of where the sense node is connected.

Transient Variations

In a worst-case transient load-dump event, the load current ramps up or down to its extreme values in a short time. Considering a positive load dump, for instance, when load current rises quickly, the LDO is at first unable to supply the load because it needs time to react and adjust (i.e., it has limited bandwidth). Filter capacitors C_{OUT}, C_B, and C_{LOAD} therefore supply this initial jump in current, most of which is derived from C_{OUT} because it presents a considerably lower impedance path than C_B and C_{LOAD} (C_{OUT} is larger than C_B and C_{LOAD} combined and impedance $(1/sC_{OUT}) + R_{ESR}$ is therefore smaller than $1/sC'_B$). The net effect is an instantaneous voltage drop across L'_{PCB} ($V_L = L'_{PCB} \cdot di_{LOAD}/dt$) that lasts as long as i_{LOAD} is changing, another instantaneous voltage drop across R_{ESR} and R'_{PCB} ($\Delta i_{LOAD} \cdot (R_{ESR} + R'_{PCB})$) and a voltage-droop response across C_{OUT}, the latter two of which last until the LDO responds and supplies the full load ($\Delta i_{LOAD}/BW \cdot C_{OUT}$). All of these parasitic effects amount to an undesired transient voltage drop in load voltage v_{LOAD}. Eventually, the IC supplies the full load and again reaches steady-state operation, at which point C_{OUT}, C_B, and C_{LOAD} cease to conduct displacement current and all these parasitic voltages disappear. For negative load dumps, similar effects occur and v_{LOAD} temporarily rises above its ideal targeted value.

1.6 Specifications

Three categories aptly describe the operating performance of a linear regulator: (1) dc- and ac-regulating (accuracy) performance, (2) power characteristics, and (3) operating requirements. The regulating performance refers to the IC's ability to *regulate* its output against variations in its operating environment. The metrics used to gauge this performance include load regulation, line regulation, power-supply rejection, temperature drift, transient load-dump variations, and dropout voltage. All these parameters essentially portray the behavior of the circuit with respect to load current, input voltage, and junction temperature. Quiescent-current flow, sleep-mode current, power efficiency, and current efficiency as well as dropout voltage, indirectly, depict the power characteristics of the regulator. Sleep-mode current is the current flowing through the IC while disabled, if an enable-disable function exists, or during low-performance mode (i.e., low-bandwidth setting). Current efficiency refers to the ratio of load to input current, which is especially important during low load-current conditions (i.e., device is idling and output current is consequently low).

The operating limits of the input voltage, output voltage, output capacitance (and associated ESR), and load current define the environment within which the regulator must operate functionally and within parametric compliance.

1.6.1 Regulating Performance

Load Regulation

Steady-state (dc) voltage variations in the output (ΔV_{OUT}) resulting from dc changes in load current (ΔI_{LOAD}) define *load regulation* (LDR) performance, which ultimately constitutes an ohmic voltage drop, that is, a linearly load-dependent voltage drop at the output of the regulator:

$$R_{LDR} = \frac{\Delta V_{OUT}}{\Delta I_{LOAD}} \equiv R_{O\text{-}REG} + R_{PAR} = \frac{R_{OL}}{1 + A_{OL}\beta_{FB}\big|_{DC}} + R'_{PCB} \approx \frac{R_{OL}}{A_{OL}\beta_{FB}\big|_{DC}} + R'_{PCB} \tag{1.11}$$

where R_{LDR} is the load-regulation resistance, $R_{O\text{-}REG}$ the closed-loop output resistance of the regulated loop, R_{OL} the open-loop output resistance from the sense node into the IC, A_{OL} the open-loop gain, β_{FB} the negative feedback-gain factor, and loop gain $A_{OL}\beta_{FB}$ is evaluated at dc because load regulation is a steady-state parameter (all frequency components are neglected). Obviously, increasing A_{OL} and decreasing PCB resistance R_{PCB}' improves load-regulation performance. Systematic input-offset voltages, which result from asymmetric currents and voltages in the feedback error amplifier, further degrade load-regulation performance. Even if the LDO were symmetric, its widely variable load would cause considerable voltage swings at internal nodes, subjecting some of the devices to asymmetric conditions. Because the open-loop gain is relatively low (the reason for this is described in a later chapter), there is a systematic load-dependent input-referred offset voltage (V_{OS_S}), which should be included in LDR:

$$R_{LDR_{EFF}} = R_{O\text{-}REG} + R'_{PCB} + \left(\frac{\Delta V_{OS\_S}}{\Delta I_{LOAD}}\right)\left(\frac{V_{OUT}}{V_{REF}}\right) = R_{O\text{-}REG} + R'_{PCB} + \left(\frac{\Delta V_{OS\_S}}{\Delta I_{LOAD}}\right)A_{CL} \tag{1.12}$$

where $R_{LDR_{EFF}}$ is the effective LDR, ΔV_{OS_S} the systematic variation of V_{OS_S} with respect to a dc change in load current ΔI_{LOAD}, and V_{OUT}/V_{REF} and A_{CL} the dc closed-loop gain from reference V_{REF} to output V_{OUT} (e.g., a 2.4 V LDO referenced with a 1.2 V bandgap has a closed-loop gain of roughly 2: 2.4 V/1.2 V).

Line Regulation

Line regulation (LNR) performance, like load regulation, is also a dc parameter and it refers to output voltage variations arising from dc changes in the input supply, in other words, to the low-frequency supply gain of the circuit (i.e., LNR is A_{IN}, which refers to $\Delta v_{OUT}/\Delta v_{IN}$). *Power-supply rejection PSR* (also known as *ripple rejection*), on the other hand, is not only the complement of supply gain A_{IN}, in that it refers to *rejection*, but it also includes the entire frequency spectrum, not just dc. Datasheets, as a result, typically quote ripple-rejection values at dc and other specific frequencies (e.g., 50 dB at dc, 40 dB at 1 kHz, etc.).

Power-supply variations affect the regulator in two ways: directly through its own supply and indirectly via supply-induced variations in reference v_{REF}. The reference, as it turns out, is a sensitive node because the regulator, being that it presents a noninverting feedback amplifier to v_{REF} (refer to Fig. 1.9 to visualize how v_{REF} and v_{OUT} relate), amplifies variations in v_{REF} by the regulator's closed-loop gain (A_{CL}). The overall supply gain A_{IN} is therefore a function of both the supply gain of the regulator $A_{IN.REG}$ and the supply gain of the reference $A_{IN.REF}$:

$$\begin{aligned} A_{IN} &\approx \frac{\Delta v_{OUT}}{\Delta v_{IN}} = A_{IN.REG} + A_{IN.REF}\, A_{CL} \\ &= \left(\frac{\Delta v_{OUT}}{\Delta v_{IN}}\right)\Bigg|_{\Delta v_{REF}=0} + \left(\frac{\Delta v_{REF}}{\Delta v_{IN}} A_{CL}\right)\Bigg|_{\Delta v_{IN.REG}=0} \end{aligned} \tag{1.13}$$

where superposition is applied and $A_{IN.REG}$ is evaluated when there is no variation in the reference and $A_{IN.REF}$ when there is no variation in the input supply of the regulator. Because PSR is A_{IN}'s complement, PSR reduces to

$$\text{PSR} \approx \frac{1}{A_{IN}} \approx \frac{1}{A_{IN.REG}} = \text{PSR}_{REG}$$

and LNR, since it only applies to the dc portion of supply gain A_{IN}, is

$$\text{LNR} = A_{IN}\big|_{DC} = A_{IN0} = \frac{1}{\text{PSR}_0} \approx \frac{1}{\text{PSR}_{REG0}} \tag{1.14}$$

Normally, v_{REF}'s supply rejection PSR_{REF} is considerably higher than PSR_{REG} so v_{REF}'s effects are often insignificant and therefore neglected, but justifying this assumption is critical.

Temperature Drift

In more generalized terms, any variation in the reference propagates to the output of the regulator through the equivalent closed-loop gain of the regulator. As such, any temperature effects on the reference, as

in the case of ripple-rejection analysis with supply-derived noise, also have adverse effects on the regulated output. These effects are in addition to the temperature effects of the regulator, which manifest themselves through temperature dependence in the input-referred offset voltage. The metric that gauges the extent of the impact of temperature on the output is *fractional temperature coefficient* (TC), which is the percentage variation of the output in response to temperature changes per degree of temperature change:

$$TC \equiv \frac{1}{V_{OUT}}\left(\frac{dv_{OUT}}{dT}\right) \approx \frac{1}{V_{OUT}}\left(\frac{\Delta v_{OUT}}{\Delta T}\right) = \frac{(\Delta v_{REF} + \Delta v_{OS})\left(\frac{V_{OUT}}{V_{REF}}\right)}{V_{OUT}\Delta T}$$
$$= \left(\frac{\Delta v_{REF} + \Delta v_{OS}}{V_{REF}}\right)\frac{1}{\Delta T} \tag{1.15}$$

where Δv_{REF} and Δv_{OS} are the temperature-induced variations of reference voltage v_{REF} and input-referred offset voltage v_{OS} and ΔT the corresponding change in temperature. Consequently, accuracy in the reference and input-offset voltage in the error amplifier are key characteristics for determining the overall temperature-drift performance of the regulator.

Transient Variations

Noise content (*ac accuracy*) is another important metric in linear regulators. The principal cause of noise is typically systematic in nature, either from a switching load or a switching supply. In the case of a switching load, the output impedance of the regulator, which is comprised of the output capacitance C_{OUT} and C_B and load-regulation output impedance r_{LDR} (ac counterpart of $R_{LDR.EFF}$) determines the extent to which the noise is suppressed. Similarly, ripple rejection limits the noise injected from the input supply. Given the mixed-signal nature of modern systems today, even with reasonable output impedances and ripple-rejection capabilities, both switching supplies and switching loads inject substantial noise into the system. Inherent shot, thermal, and $1/f$ noise are normally not as important and often neglected in specifications as a result. This is not to say, however, systematic switching noise *always* overwhelms inherent noise, especially when considering extremely sensitive and high-performance applications.

Transiently induced voltage variations also contribute to the overall ac accuracy of the regulator. The worst-case variation occurs when the load current suddenly transitions from its lowest rated value (e.g., zero) to its maximum peak, or vice versa, which comprise the positive and negative *load-dump* conditions graphically illustrated in Fig. 1.12. The response is normally asymmetrical in nature because

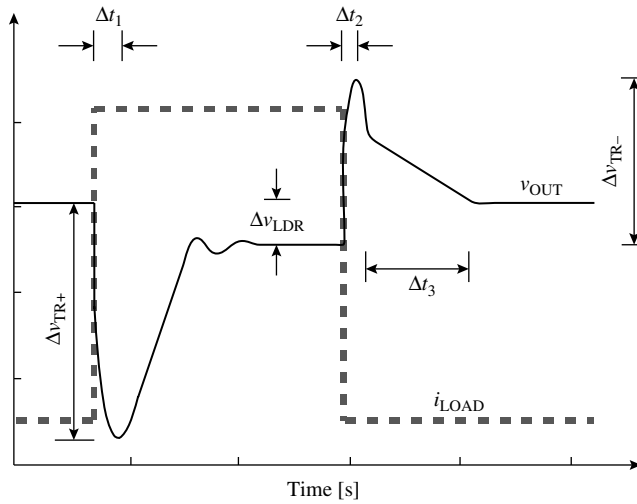


Figure 1.12 Typical transient response to positive and negative load dumps (i.e., sudden load-current changes).

the error amplifier's ability to charge and discharge the input of the pass device is also asymmetrical. To be more specific the output stage of the error amplifier is a buffer whose purpose is to drive the input of the pass device, which is highly capacitive, given its large physical dimensions. For simplicity, the buffer normally conforms to class-A operation, which is capable of pushing *or* pulling substantial current, but not both. As a result, the buffer typically slews into the highly capacitive node in one direction and not the other, giving rise to different response times and therefore asymmetrical effects on the regulated output. A symmetrical response is possible with class-B and class-AB buffers but they usually require more silicon real estate, complexity, and noise, which amount to higher cost.

In the case of a positive load dump, when current suddenly rises, the additional load current (Δi_{LOAD}) discharges the output filter capacitors until enough time elapses to allow the loop (i.e., the regulator) to respond (in Δt_1). The internal slew-rate conditions of the feedback loop (i.e., when class-A buffer slews) normally govern the extent of response time Δt_1, which exceeds the corresponding bandwidth time ($1/BW_{CL}$). After the circuit has time to react, the pass device responds by supplying load current i_{LOAD} and additional current to charge and slew the output filter capacitors back to their targeted regulated voltage. The regulated output voltage ultimately settles to the voltage corresponding to the new load-current value, which is the ideal voltage minus the load-regulation effect of the loop (ΔV_{LDR}). When a negative load dump occurs (i.e., current falls quickly), the extra current the pass device initially sources, which cannot decrease until the

regulator has time to react, charges the output filter capacitors. When enough time elapses to allow the loop to respond ($\Delta t_2 \approx 1/\mathrm{BW}_{\mathrm{CL}}$), the regulator stops sourcing current and allows whatever sinking capabilities it has (i.e., feedback network) to discharge and slew the output capacitor back to its targeted output voltage.

The resulting variation of the regulated output in response to these load dumps degrades the overall regulating performance of the regulator and amounts to *transient accuracy*. As noted in Fig. 1.12, the regulator ultimately experiences load-regulation effects ($\Delta V_{\mathrm{LDR}} = \Delta I_{\mathrm{LOAD}} R_{\mathrm{LDR.EFF}}$), which, for technical accuracy, should be distinct and decoupled from the total effects transient-response accuracy has on the output (Δv_{TR}). Load-regulation voltage drop ΔV_{LDR} is therefore subtracted from total load-dump-induced variations $\Delta v_{\mathrm{TR+}}$ and $\Delta v_{\mathrm{TR-}}$.

During a positive load dump, when load current suddenly increases, referring to the effective load model shown in Fig. 1.11, output capacitor C_{OUT} and total bypass capacitor C_B' supply additional load current Δi_{LOAD}. Since C_{OUT} is normally more than an order of magnitude higher than C_B' (impedance $1/\mathrm{s}C_{\mathrm{OUT}}$ is much smaller than $1/\mathrm{s}C_B'$), most of Δi_{LOAD} flows through C_{OUT} and ESR resistor R_{ESR}, the result of which is an instantaneous voltage across R_{ESR} equivalent to

$$\Delta v_{\mathrm{ESR}} \approx \left(\frac{C_{\mathrm{OUT}}}{C_{\mathrm{OUT}} + C_B'} \right) \Delta i_{\mathrm{LOAD}} R_{\mathrm{ESR}} \tag{1.16}$$

and a droop voltage across C_{OUT} and C_B'. The aggregate effect on the output, considering this condition persists until the regulator reacts (i.e., Δt_1 after the onset of the load dump), is

$$\begin{aligned} \Delta v_{\mathrm{TR+}} &\approx \left(\frac{\Delta i_{\mathrm{LOAD}}}{C_{\mathrm{OUT}} + C_B'} \right) \Delta t_1 \; + \Delta v_{\mathrm{ESR}} \\ &\approx \left(\frac{\Delta i_{\mathrm{LOAD}}}{C_{\mathrm{OUT}} + C_B'} \right) \left(\Delta t_1 \; + C_{\mathrm{OUT}} \, R_{\mathrm{ESR}} \right) \end{aligned} \tag{1.17}$$

The bypass capacitors have a filtering effect on this total variation, decreasing its magnitude when more low-ESR capacitance is present.

As already mentioned, the slewing conditions of the class-A buffer against the parasitic capacitance presented by the large series pass device (C_{PAR}) set response time Δt_1. Before the onset of these slew-rate conditions, however, the loop must have sufficient time to react and initiate the slew command, producing in the process a bandwidth-limited delay that is proportional to the reciprocal of the closed-loop bandwidth ($\mathrm{BW}_{\mathrm{CL}}$) of the regulator (approximately $0.37/\mathrm{BW}_{\mathrm{CL}}$—refer to App. A). After this, the biasing current of the buffer (I_{BUF}) slews parasitic capacitor C_{PAR} (producing Δv_{PAR}) until enough drive exists

across the power pass device to fully supply load current i_{LOAD}. Total response time Δt_1 therefore approximates to

$$\Delta t_1 \approx \frac{0.37}{BW_{CL}} + C_{PAR}\left(\frac{\Delta v_{PAR}}{I_{BUF}}\right) \approx \frac{0.37}{BW_{CL}} + C_{PAR}\left(\frac{\Delta i_{LOAD}}{g_{mp} I_{BUF}}\right) \tag{1.18}$$

where g_{mp} is the transconductance of the power pass device. If slew-rate current I_{BUF} is sufficiently high, response time Δt_1 reduces to $0.37/BW_{CL}$. This response-time improvement and independence to C_{PAR} result at the expense of quiescent-current flow, that is to say, reduced power efficiency and battery life.

During a negative load dump, when load current suddenly drops, the regulator continues to source initial load current I_{LOAD}, pushing the difference (Δi_{LOAD}) into output filter capacitors C_{OUT} and C_B'. As a result, similar to a positive load dump, R_{ESR} incurs an instantaneous voltage across its terminals and output filter capacitors C_{OUT} and C_B' slew until the regulator responds, at which point the pass device shuts off and allows the output to drop to its steady-state value. Assuming the class-A buffer slews for positive load dumps and not negative load dumps, only the closed-loop bandwidth of the loop sets the corresponding negative load-dump response time to approximately $0.37/BW_{CL}$, giving rise to a full negative load-dump variation (Δv_{TR-}) of

$$\Delta v_{TR-} \approx \left(\frac{\Delta i_{LOAD}}{C_{OUT} + C_B'}\right)\Delta t_2 \; + \Delta v_{ESR} \approx \left(\frac{\Delta i_{LOAD}}{C_{OUT} + C_B'}\right)\left(\frac{0.37}{BW_{CL}} + C_{OUT} \; R_{ESR}\right) \tag{1.19}$$

To decouple load regulation from transient-accuracy performance, LDR voltage ΔV_{LDR} is subtracted from positive and negative transient variations Δv_{TR+} and Δv_{TR-}. Transient accuracy therefore summarizes to

$$\Delta v_{TR} \approx \begin{cases} + \left[\left(\dfrac{\Delta i_{LOAD}}{C_{OUT} + C_B'}\right)\left(\dfrac{0.37}{BW_{CL}} + C_{OUT} R_{ESR}\right) - \Delta i_{LOAD} \; R_{LDR_{EFF}}\right] \\ - \left[\left(\dfrac{\Delta i_{LOAD}}{C_{OUT} + C_B'}\right)\left(\dfrac{0.37}{BW_{CL}} + C_{PAR}\left(\dfrac{\Delta i_{LOAD}}{g_{mp} I_{BUF}}\right) + C_{OUT} \; R_{ESR}\right) - \Delta i_{LOAD} \; R_{LDR_{EFF}}\right] \end{cases} \tag{1.20}$$

which is often simplified to

$$\begin{aligned} \Delta V_{TR(max)} &\approx \text{Max}\left(\Delta V_{TR+}, \Delta V_{TR-}\right) \\ &\approx \left(\frac{\Delta i_{LOAD}}{C_{OUT} + C_B'}\right)\left(\frac{0.37}{BW_{CL}} + C_{PAR}\left(\frac{\Delta i_{LOAD}}{g_{mp} i_{BUF}}\right) + C_{OUT} \; R_{ESR}\right) \\ &\quad - \Delta i_{LOAD} \; R_{LDR_{EFF}} \end{aligned} \tag{1.21}$$

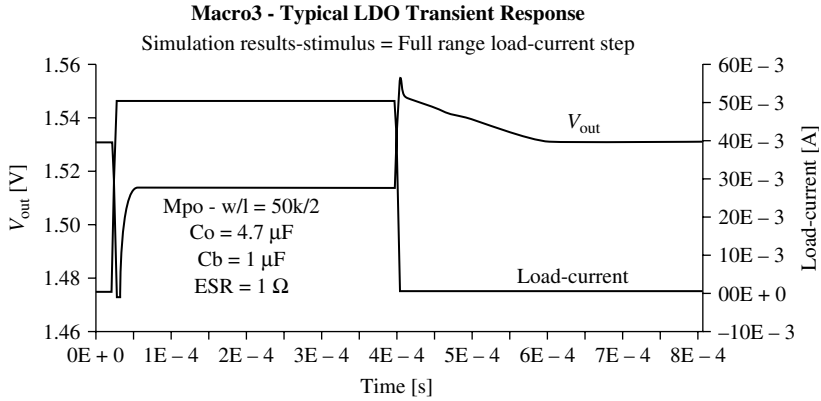


FIGURE 1.13 Simulated transient-response performance of an LDO under load-dump conditions.

where the slew-rate-limited buffer response is worse. Figure 1.13 shows the simulation results of a 1.53-V LDO under positive and negative 50 mA load dumps and in the presence of a 4.7 µF electrolytic capacitor with 1 Ω of ESR in combination with an equivalent bypass capacitance of 1 µF at the output. The LDO is comprised of an ideal voltage reference and an amplifier macromodel driving a 50 mm/2 µm PMOS power transistor (Mpo). The results show the output suffers an LDR voltage drop of 20 mV (Δv_{LDR}) and a worst-case transient voltage drop of 55 mV (Δv_{TR+}), translating to a transient accuracy of 2.3% (or 35 mV/1.53 V) and load-regulation accuracy of 1.3% (or 20 mV/1.53 V).

As the example shows, transient accuracy is a significant portion of the overall performance of the regulator, which is why a 3–8% margin is often allocated to this phenomenon. Using high-frequency capacitors, which have low ESRs and cost more money, considerably reduces this variation. Increasing the response time of the regulator has similar benefits, but at the expense of more quiescent current. Increasing the rise and fall times of the load dumps, in other words, relaxing them to the point of exceeding the response time of the loop practically negates all adverse transient effects, and this is how many data books specify it, when rise and fall times are on the order of microseconds. However, in state-of-the-art applications, the load is often synchronized to digital-signal-processing (DSP) and microprocessor (µP) clocks with frequencies ranging from several megahertz to gigahertz, subjecting the regulator to substantially quick load dumps (in the nanosecond region), which is where transient accuracy is worse. It is therefore important to determine and design for a practical and realistic load; fully understanding the load allows the designer to better trade accuracy performance for silicon real estate and power efficiency.

Example 1.1 Determine the load-regulation and transient accuracy of a 100 mA 1.5 V regulator with 500 kHz bandwidth and 300 mΩ equivalent load-regulation resistance when subjected to 0–100 mA load dumps. Assume the output filter is comprised of a 10 µF capacitor with an ESR of 0.2 Ω and an equivalent bypass capacitance of 1 µF. The parasitic input capacitance of the power pass device is roughly 100 pF and its transconductance 300 mA/V, and it is driven by a class-A buffer with a constant pull-down current of 5 µA.

$$\Delta V_{\text{LDR}} \approx \Delta I_{\text{LOAD}}\, R_{\text{LDR}_{\text{EFF}}} \approx 100\text{ mA} \cdot 300\text{ m}\Omega \approx 30\text{ mV}$$

$$\therefore \quad \text{LDR Accuracy} = \frac{\Delta V_{\text{LDR}}}{V_{\text{OUT}}} \approx \frac{30\text{ mV}}{1.5\text{ V}} \approx 2.0\%$$

$$\Delta V_{\text{TR}} \approx \left(\frac{\Delta i_{\text{LOAD}}}{C_{\text{OUT}} + C_B'}\right)\left(\frac{0.37}{\text{BW}_{\text{CL}}} + C_{\text{PAR}}\left(\frac{\Delta i_{\text{LOAD}}}{g_{\text{mp}}\, I_{\text{BUF}}}\right) + C_{\text{OUT}}\, R_{\text{ESR}}\right) - \Delta V_{\text{LDR}}$$

$$\frac{100\text{mA}}{10\mu\text{F} + 1\mu\text{F}}\left(\frac{0.37 \cdot 2\pi}{500\text{kHz}} + 100\text{pF}\frac{100\text{mA}}{300\frac{\text{mA}}{V} \cdot 5\mu\text{A}} + 10\mu\text{F} \cdot 0.2\,\Omega\right) - 30\text{mV} \approx 91\text{mV}$$

$$\therefore \quad \text{Transient Accuracy} = \frac{\Delta v_{\text{TR}}}{V_O} \approx \frac{91\text{mV}}{1.5\text{V}} \approx 6.1\%$$

Gain Error

The effects of dc and transient loads, line (supply voltage), and temperature variations on the regulated output amount to dc and ac accuracy. Gain error G_E in the feedback loop, however, which is worse for lower loop gains, also has a negative impact on accuracy performance:

$$G_E \approx \frac{A_{\text{CL}} - A_{\text{CL\_I}}}{A_{\text{CL\_I}}} = \left(\frac{A_{\text{OL}}}{1 + A_{\text{OL}}\,\beta_{\text{FB}}} - \frac{1}{\beta_{\text{FB}}}\right)\beta_{\text{FB}} = \frac{-1}{1 + A_{\text{OL}}\,\beta_{\text{FB}}} \qquad (1.22)$$

where A_{CL} is the closed-loop forward gain from the reference to the output, β_{FB} the feedback-gain factor, and $A_{\text{CL\_I}}$ and $1/\beta_{\text{FB}}$ the ideal closed-loop gain (equal to one if v_{OUT} is meant to be regulated to V_{REF}). Given a gain error from V_{REF} to v_{OUT}, the resulting systematic variation in the output (ΔV_{G_E}) simplifies to the product of reference V_{REF} and gain error G_E and always results in an output voltage that is below its target (note the negative sign in the G_E relationship)

$$\Delta V_{G_E} = V_{\text{REF}}\, G_E \qquad (1.23)$$

Overall Accuracy

Ultimately, accuracy refers to the total output voltage variation, which includes both systematic ($\Delta v_{\text{Systematic}}$) and random ($\Delta v^*_{\text{Random}}$) components. Systematic offsets are consistent, monotonic, and considering their effects are often in the millivolt region, mostly linear. The combined impact of several systematic offsets on the output is therefore the linear sum of their individual effects. For instance, if a 1–50 mA increase in load current pulls v_{OUT} 20 mV below its target under a 5 V supply and a 5–3 V decrease in supply pulls v_{OUT} 5 mV when loaded with 1 mA, their combined impact on v_{OUT} is approximately a 25 mV drop, assuming a linear relationship. Systematic offsets have a polarity and will sometimes cancel one another, depending on the circuit and its application. Random offsets, on the other hand, are neither consistent nor monotonic, and their effects must consequently be treated under probabilistic terms, where the combined effects of several components is the root-square sum of its constituent factors:

$$\Delta v^*_{\text{Random}} \approx \sqrt{\sum_{1}^{N} \Delta v_1^{*2}} \tag{1.24}$$

where N denotes the total number of random effects on the output. In the end, the total variation of the output exhibits combined systematic and random components and its accuracy ultimately conforms to similar terms:

$$\text{Accuracy} \approx \frac{\sum \Delta v_{\text{Systematic}} \pm \Delta v^*_{\text{Random}}}{V_{\text{OUT}}} \tag{1.25}$$

Load, line, transient, temperature, and gain effects are all systematic, monotonic, and for the most part, linear, whereas threshold, transconductance parameter, and reverse-saturation mismatch offsets and process-induced variations in the reference are all random and probabilistic. The total accuracy performance of a voltage regulator is consequently

$$\text{Accuracy} \approx \frac{\Delta V_{\text{LDR}} + \Delta V_{\text{LNR}} + \Delta v_{\text{TR}} + \Delta V_{\text{TC}} + \Delta v_{\text{REF}}\left(\dfrac{V_{\text{OUT}}}{V_{\text{REF}}}\right) + V_{\text{REF}}\, G_E}{V_{\text{OUT}}}$$

$$\pm \frac{\sqrt{\left(\Delta v^*_{\text{REF}}\left(\dfrac{V_{\text{OUT}}}{V_{\text{REF}}}\right)\right)^2 + \left(v^*_{\text{OS}}\left(\dfrac{V_{\text{OUT}}}{V_{\text{REF}}}\right)\right)^2 + \left(V_{\text{REF}}\, G^*_E\right)^2}}{V_{\text{OUT}}} \tag{1.26}$$

where ΔV_{LDR}, ΔV_{LNR}, Δv_{TR}, Δv_{TC}, and Δv_{REF} are systematic output voltage variations resulting from load regulation, line regulation, transient load dumps, temperature drift, and systematic variations in reference v_{REF}; V_{OUT}/V_{REF} the ideal closed-loop gain; and Δv_{REF}^*, v_{OS}^*, and G_E^* the random process-induced variations in v_{REF}, error amplifier's input-referred offset, and gain variations, respectively. Adhering strictly to a linear summation of all terms, systematic and random, produces the absolute worst-case performance for all devices at all times, which is unrealistically pessimistic. A strict probabilistic sum (i.e., root sum of squares), on the other hand, is unrealistically optimistic because the consistent and repeatable nature of the systematic components is underestimated.

In practice, transient load-dump effects dominate accuracy, but like other effects, though for different reasons, overall accuracy specifications often exclude them. The fact is these effects depend on how fast the load can possibly rise or fall (di_{LOAD}/dt); in other words, they depend strongly on the application. Line- and load-regulation effects in v_{REF} are also excluded, but unlike their transient counterpart, their impact on v_{REF}, when compared to other factors, is minimal. Package-stress effects and other process-induced random variations on v_{REF} are normally so severe that temperature-drift performance is often difficult to discern, which is why all systematic variations in the reference (Δv_{REF}) are often altogether absorbed by Δv_{REF}^*. Gain error G_E and process-induced variations in G_E are similarly neglected because their impact is relatively low. The random effects of offset voltage v_{OS}^* and reference voltage variation Δv_{REF}^* are often combined into one random-variation parameter ($\Delta v_{REF.OS}^*$) because they are characterized together—measuring v_{OS}^* alone is not as useful and requires more work. Ultimately, accuracy tends to simplify to

$$\text{Accuracy} \approx \frac{\Delta V_{LDR} + \Delta V_{LNR} + \Delta V_{TC} \pm \sqrt{\left(\Delta v^*_{REF.OS} \frac{V_{OUT}}{V_{REF}}\right)^2}}{V_{OUT}} \quad (1.27)$$

Quoted accuracies for linear regulators typically fall in the 1–3% range, and that excludes transient performance, which can take another 1–7%, depending on the loading profile presented by the application at hand.

Example 1.2 Determine the total accuracy performance of the 2.4 V regulator shown in Fig. 1.14. Assume its forward open-loop gain is 60 dB and output voltage variations with their corresponding three-sigma variations due to temperature, load-current, input-supply, full-range transient load-current changes are 3 ± 30 mV, 20 ± 5 mV, 8 ± 2.5 mV, and 140 ± 5 mV, respectively.

Note the temperature coefficient of the device, which is 3 ± 30 mV divided by the product of 2.4 V and 125°C, is not consistent across process because

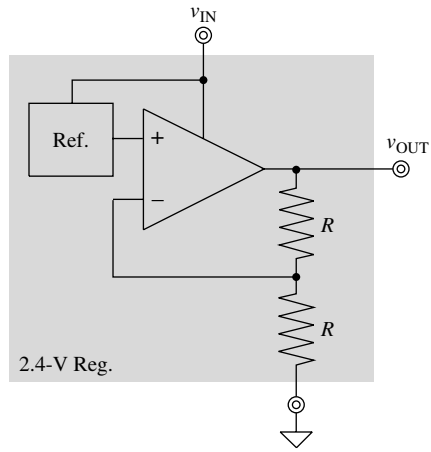


FIGURE 1.14
Three-terminal 2.4-V linear regulator.

random effects overwhelm systematic effects. This is typical in plastic-packaged, temperature-compensated voltage references, which tend to suffer from pronounced random-induced offset variations. This randomness results because the fillers in the package used to increase its reliability have random shapes and therefore exert random pressure across the die, resulting in random piezoelectric effects in the reference.

$$\text{Ideal closed-loop gain } A_{CL-I} \approx \frac{V_{OUT}}{V_{REF}} = \frac{1}{\beta_{FB}} = \frac{\mathrm{R}+\mathrm{R}}{\mathrm{R}} = 2$$

$$\therefore \qquad \text{Reference } V_{REF} = \frac{2.4\text{ V}}{2} = 1.2\text{ V}$$

$$\text{And gain error } G_E = \frac{-1}{1+A_{OL}\beta_{FB}} = \frac{-1}{1+\dfrac{1000}{2}} = -0.002$$

Transient variation Δv_{TR} is the total output variation in response to a transient load dump minus its load-regulation contribution ΔV_{LDR}:

$$\therefore \qquad \Delta v_{TR} = 140\text{ mV} - 20\text{ mV} = 120\text{ mV}$$

$$\text{And} \qquad \pm 5\text{ mV} = \sqrt{\Delta v_{TR}^{*\,2} + \Delta V_{LDR}^{*\,2}} = \sqrt{\Delta v_{TR}^{*\,2} + (5\text{ mV})^2}$$

Δv_{TR}^{*} is negligible

And since the regulation performance of the entire circuit was quoted, with v_{REF} as part of the system, all systematic and random variations in the reference

(Δv_{REF} and Δv_{REF}^{*}) are already included in load- and line-regulation effects Δv_{LDR} and Δv_{LNR}; in other words, for all practical purposes, $\Delta v_{REF} = 0$ and $\Delta v_{REF}^{*} = 0$.

$$\text{Accuracy} \approx \frac{\Sigma \Delta V_{\text{Systematic}} \pm \sqrt{\Sigma \Delta V_{\text{Random}}^{*2}}}{V_{OUT}}$$

$$\approx \frac{\Delta V_{LDR} + \Delta V_{LNR} + \Delta v_{TR} + \Delta V_{TC} + V_{REF}\, G_E}{V_{OUT}}$$

$$\pm \frac{\sqrt{\Delta V_{TC}^{*2} + \Delta V_{LDR}^{*2} + \Delta V_{LNR}^{*2}}}{V_{OUT}}$$

$$\approx \frac{20\text{ mV} + 8\text{ mV} + 120\text{ mV} + 3\text{ mV} + 1.2\text{ V} \cdot 2\text{ m}}{2.4\text{ V}}$$

$$\pm \frac{\sqrt{(30\text{ mV})^2 + (5\text{ mV})^2 + (2.5\text{ mV})^2}}{2.4\text{ V}}$$

$$\approx 6.4\% \pm 1.3\% \leq 7.7\%$$

Or, if transient accuracy is excluded,

$$\text{Accuracy} \approx 1.4\% \pm 1.3\% \leq 2.7\%.$$

The regulator is consequently said to have an accuracy performance of 2.7%.

1.6.2 Power Characteristics

The basic functions of a voltage regulator are to *condition power* and *transfer energy* from a source to an electronic load. Accuracy, as a metric for regulation performance, gauges the power-conditioning capabilities of a regulator whereas *efficiency* relays information about energy and its ability to transfer it. Ideally, a regulator transfers only the energy the load demands from the source to the load, but in practice, energy is always lost in the process of conditioning power, which is why efficiency, defined as the ratio of delivered energy E_{LOAD} to stored energy E_{SOURCE}, gauges what fraction of the source energy actually reaches the load. Since energy is simply power P over the span of time t (i.e., $E = \int P \cdot t$), assuming power remains constant throughout time, *energy efficiency* and *power efficiency* (η) amount to the same:

$$\eta \approx \frac{E_{LOAD}}{E_{SOURCE}} = \frac{P_{LOAD} \cdot t}{P_{SOURCE} \cdot t} = \frac{P_{LOAD}}{P_{SOURCE}} = \frac{I_{LOAD} V_{OUT}}{I_{IN} V_{IN}} \qquad (1.28)$$

where V_{IN} and I_{IN} are the steady-state input supply voltage and its associated current. Efficiency is an extremely important parameter in the world of portable electronics because it determines operational life, given its source is either charge- and/or fuel-constrained, as in the case of battery-powered devices.

In transferring energy, a regulator loses power across its power-conducing series device and through its control loop. The power dissipated by the series pass switch is the product of the voltage across it (i.e., input-voltage difference $V_{IN} - V_{OUT}$) and its current (I_{LOAD}). Similarly, the power dissipated by the regulating loop is the product of the quiescent current (I_Q) the loop requires to function and the voltage across its rails (i.e., $P_Q = V_{IN}\, I_Q$). As a result, efficiency expands and simplifies to

$$\eta = \frac{V_{OUT}\, I_{LOAD}}{V_{OUT}\, I_{LOAD} + (V_{IN} - V_{OUT})I_{LOAD} + V_{IN}\, I_Q}$$

$$= \frac{V_{OUT}\, I_{LOAD}}{V_{IN}(I_{LOAD} + I_Q)} \approx \frac{V_{OUT}\, \eta_I}{V_{IN}} \tag{1.29}$$

where η_I is *current efficiency*, which is the ratio of I_{LOAD} to total input current I_{IN} (or $I_{LOAD} + I_Q$):

$$\eta_I \approx \frac{I_{LOAD}}{I_{IN}} = \frac{I_{LOAD}}{I_{LOAD} + I_Q} \tag{1.30}$$

A system designer therefore assigns as low a supply voltage a regulator can handle (i.e., low V_{IN}) while meeting its regulation objectives. The IC designer must design, as a result, a circuit that not only sustains the lowest possible voltage across its power pass device (without pushing it into dropout to continue enjoying the benefits of regulation) but also achieves the highest possible current efficiency (i.e., lowest I_Q during light loading conditions).

Dropout voltage is often the limiting factor in efficiency performance. From a specification standpoint, the minimum voltage sustained across the pass device when the circuit ceases to regulate is the dropout voltage. The loop, during this condition, drives the pass device to source maximum current to the output, but has no ac gain to offer and therefore no regulation capabilities. However, if V_{DO} is low and the regulator biased slightly above its dropout region, that is, V_{OUT} is slightly greater than $V_{IN} - V_{DO}$, optimum efficiency performance is achieved:

$$\eta = \frac{v_{OUT}\, \eta_I}{v_{IN}} \le \frac{(V_{IN} - V_{DO})\eta_I}{V_{IN}} \le \left(1 - \frac{V_{DO}}{V_{IN}}\right) \tag{1.31}$$

Since the loop is "railed out" and the power pass device is maximally driven during dropout conditions, V_{DO} is an ohmic drop, which means V_{DO} and equivalent switch-on resistance R_{ON} describe the same effect and are linked by I_{LOAD} as

$$V_{DO} = I_{LOAD}\ R_{ON} \tag{1.32}$$

The total switch-on resistance is comprised of all series resistors from the regulation node (i.e., sense point) to input supply voltage V_{IN}, which normally amounts to the on-resistance of the intrinsic power device (R_{ON_I}) and the series parasitic resistances associated with the input and output bond wires (R_{BW}) and any metallization (R_{METAL}) used to link them,

$$R_{ON} = R_{BW} + R_{ON\_I} + R_{METAL} \tag{1.33}$$

Reasonable dropout voltages for portable applications are on the order of 200–300 mV.

Example 1.3 Determine the worst- and best-case full- and zero-load efficiency performance of a 0–100 mA 1.2 V linear regulator with an extrinsic on-resistance of 500 mΩ and quiescent dc current of 10 μA. The regulator draws its power from a 0.9–1.6 V NiCd battery.

Worst-case full-load (i.e., I_{LOAD} is at its maximum value) efficiency occurs when V_{IN} is at its maximum:

$$\eta = \frac{V_{OUT}\ I_{LOAD\_MAX}}{V_{IN\_MAX}(I_{LOAD\_MAX} + I_Q)} = \frac{1.2 \cdot 100\ \text{m}}{1.6 \cdot (100\ \text{m} + 10\ \mu)} \approx \frac{1.2}{1.6} = 75\%$$

Best-case full-load efficiency occurs when I_{LOAD} is at its maximum point and the device is operated slightly above its dropout region:

$$V_{DO} = I_{LOAD\_MAX}\ R_{ON} = 100\ \text{m} \cdot 500\ \text{m} = 50\ \text{mV}$$

$$\therefore \quad \eta = \frac{V_{OUT} I_{LOAD\_MAX}}{V_{IN}(I_{LOAD\_MAX} + I_Q)} \approx \frac{V_{OUT} I_{LOAD\_MAX}}{(V_{OUT} + V_{DO}) \cdot (I_{LOAD\_MAX} + I_Q)}$$

$$= \frac{1.2 \cdot 100\ \text{m}}{(1.2 + 50\ \text{m}) \cdot (100\ \text{m} + 10\ \mu)} \approx \frac{1.2}{1.25} = 96\%$$

Since the output power is zero during zero-load current conditions, efficiency is also zero, irrespective of other operating conditions:

$$\eta = \frac{V_{OUT}\ I_{LOAD\_MIN}}{V_{IN}(I_{LOAD\_MIN} + I_Q)} = \frac{0}{V_{IN}\ I_Q} = 0\%$$

which is why quiescent current alone is an important parameter in portable devices, since load current often falls to zero or low-current levels.

1.6.3 Operating Environment

Input Supply

The operating range of *input voltage* V_{IN}, *output voltage* V_{OUT}, *output capacitor* C_{OUT}, capacitor *ESR* R_{ESR}, and *load current* I_{LOAD} describes the working limits of the regulator. While the application determines the minimum acceptable range, the circuit sets the outer boundaries. For instance, while a Li Ion normally spans the range of 2.7–4.2 V, the breakdown voltages of the process technology used to build the circuit sets the upper boundary to, say, 6 V and the minimum headroom requirements of the circuit, the lower limit, to maybe 2.2 V. As a result, choosing the process in which to build the circuit is one of the first design tasks, and its primary goal is normally for the breakdown voltages to exceed the maximum supply-voltage limit of the application. After that, for optimum speed and efficiency, one process may be picked over the others because of its cost and the components it offers. The designer is then tasked to design the circuit such that its headroom limits are low enough to sustain the application.

If a switching converter supplies power to the linear regulator, the system designer sets the optimal dc target value for V_{IN}. In this case, the specified V_{IN} represents a combination of several factors, the most important two of which are the headroom limits of the linear regulator and its voltage drop ($V_{IN} - V_{OUT}$), which is important for efficiency. The headroom limits of the load and the overall regulator efficiency similarly set output voltage V_{OUT}. Another difference in converter-supplied applications is the need for start-up control. The switching converter requires time to ramp up and the overall system must stay in control and within safety limits (e.g., power switches kept within their rated power limits) throughout the start-up period. To remain in control, however, certain functions must operate properly during start-up, before the supplies reach their parametric targets, which is why low-headroom circuits (i.e., circuits that tolerate low V_{IN} values) are appealing in power-management applications.

Compensation

Stability is another parameter to consider and, along with transient accuracy, sets the acceptable range for output capacitance C_{OUT} and resistance R_{ESR}. If the dominant low-frequency pole is at the output, C_{OUT} must exceed a certain value to guarantee stability. Similarly, given load-dump demands, C_{OUT} must exceed a minimum target to prevent the output from drooping below or above its specified accuracy limits. If the dominant low-frequency pole is inside the circuit, on the other hand, C_{OUT} must stay below a maximum value to ensure the frequency location of the output pole is not pulled low enough to compromise stability conditions. Transient-response and supply-ripple rejection requirements demand low ESR values. Stability, on the other

hand, may benefit or suffer from the presence of an ESR, which is why a range is often specified, not just a lower or an upper boundary limit.

Load Current

Although the application normally sets the load-current range, the circuit, as in the case of V_{IN}, imposes outer limits. First, the current-voltage-temperature power-rating limits, otherwise known as the *safe-operating area* (SOA), of the series pass device set "hard" upper-boundary limits. Additionally, a wide load range also implies the ac operating conditions of the regulator change considerably, making it difficult to guarantee stability. The output resistance of the series pass device, for instance, which has a significant impact on frequency response when not in dropout, is strongly dependent on I_{LOAD}. Twenty to thirty years ago, applications only demanded maybe two to three decades of change in load current (e.g., 1 mA to 1 A) because wall outlets, which offer virtually boundless energy when compared to batteries, supplied them. In today's portable market, however, battery-drain current, in an effort to maximize lifetime, cannot remain relatively high indefinitely and must therefore conform to power-moded schemes where the load may be completely (or almost completely) shut off. In such cases, load current may span eight to nine decades, if not more, and have a considerable impact on stability and consequently on load-current range. System and IC designers must challenge and fully justify the load-current range a system demands before unnecessarily designing for impossible conditions.

1.7 Simulations

1.7.1 Functionality

The role of simulations in state-of-the-art designs is increasingly important. Their general objectives are to (1) verify functionality and (2) ascertain parametric-compliance limits. As in programming, however, they are as good as their inputs: "*garbage in, garbage out*." Designers should therefore simulate only when they *think* they know what to expect so that they may properly evaluate the simulation. If the results do not conform to expectations, there is either a problem with the simulation itself or the circuit, so the first thing to do is ensure the operating conditions, models, and so on of the simulation are correct, and if so, a reevaluation of the circuit (without the computer) is in order. This process is iterated until the results match expectations, at which point the designer is in a better position to make important design choices and performance tradeoffs.

1.7.2 Parametric Limits

The second important objective of simulations is to ascertain the parametric limits of the circuit when subjected to extreme process-corner variations and operating conditions, oftentimes referred as process-voltage-temperature (PVT) corners, even though extreme load current, output capacitance, and other operating conditions are also considered. These simulations ascertain systematic and process-wide variations of all performance parameters by repeating each circuit test and successively changing models and operating conditions to include all possible combinations. For instance, weak NMOS, strong PMOS, and nominal NPN transistors, along with high resistor and capacitor values, may constitute one model set of many similar and distinct model combinations. Similarly, high temperature, low input voltage, low output capacitance, high ESR, and so on also comprise a condition set to which the circuit will be exposed while performing worst-case corner simulations.

Process engineers often guarantee parameters that perform better than what they claim because their intent is to increase die yield (i.e., profits for the company), which means process-corner simulations is, on probabilistic terms, pessimistic: an exhaustive linear combination of six-sigma variables is unrealistic. The caveat, however, from a designer's perspective, is that an analog circuit has an infinite number of operating conditions (e.g., start-up conditions subject a circuit to an infinite number of bias points—large-signal behavior—and therefore an infinite set of ac conditions) and it is impossible to simulate them all, one at a time. This is why a linear sum of six-sigma process variations, in combination with good engineering judgment, is believed to mitigate the risk associated with analog IC design and help justify the cost of fabrication. Fast product-development cycles also rely on these extreme corner scenarios to increase the chances of building a sufficiently robust prototype to meet all parametric limits after only one fabrication cycle, achieving the coveted *first-pass success*.

1.8 Summary

The purpose of a regulator is to *regulate* the output voltage against all possible operating conditions, from load and supply to temperature variations. They differ from references only in that they supply load current, but this seemingly insignificant fact adds considerable complexity and challenges to the problem. However, linear regulators, when compared with switching regulators, are simpler, faster, and less noisy, but they suffer from limited power efficiency, which is why low dropout voltages are so appealing in linear regulators for portable, battery-powered applications where single-charge battery life is crucial. Consequently, among other categories, like power level and frequency-compensation strategy, dropout voltage is an important

metric in linear regulators. In the end, however, irrespective of class, all regulators must ultimately conform to the accuracy, power, and operating conditions a given application demands, defining in the process, among others, the load-regulation, line-regulation, temperature-drift, and ripple-rejection performance requirements of the regulator. In designing for these performance objectives under all possible conditions, simulations are extremely useful, but only when the designer knows (or *thinks* he or she knows) what to expect, as simulations are as misleading as their programming variables. To be able to anticipate circuit response, the designer must therefore understand the basics of analog circuit design, especially as it pertains to linear regulators, which is the focus of the next few chapters.

CHAPTER 2
Microelectronic Devices

Now that Chap. 1 sets the context and performance objectives for linear regulators, it is important to describe the foundation on which an analog IC designer relies to build a circuit. The purpose is to present, review, and highlight those aspects of analog IC design that are critical in designing linear regulators and other analog and mixed-signal systems. The next couple of chapters therefore focus on the basic building blocks used to build analog circuits, from devices, current mirrors, single-transistor amplifiers, and differential pairs to voltage and current references and feedback circuits, each of which plays a critical role in a linear regulator, as in most other analog circuits. Ultimately, their combined large- and small-signal response determines the extent to which a system is able to achieve its overall performance objectives, which is why understanding their individual effects first is critical. Subsequent chapters will then use and combine these basic analog building blocks to design and analyze the operating limits of linear regulator circuits.

2.1 Passives

2.1.1 Resistors

The most basic, yet most powerful device an IC designer uses is a *resistor*. Although not frequently thought this way, it is most often used as a short-circuit link to conduct an electrical signal from one circuit to another. In fact, any material used to connect two points, be it aluminum, polysilicon, or doped silicon, introduces resistance to the flow of charge carriers, demanding power and energy. These links often affect circuit performance by introducing parasitic ohmic voltages (v_R) and power losses (P_R):

$$v_R = i_R R \tag{2.1}$$

and

$$P_R = i_R v_R \tag{2.2}$$

where i_R is the current flowing through the device and R its resistance. *Every node in a circuit will therefore introduce resistance, demand power, and use energy*, which is why design engineers use metal or heavily doped polysilicon materials for this purpose, to ensure their resistances are negligibly small (e.g., milliohms) when compared to the resistances of the devices connected to them (e.g., kilohms to megohms).

In a more traditional sense, however, resistors convert currents into voltages, and when used for this purpose, resistance is no longer parasitic but by design. Generally, referring to Fig. 2.1, increasing the length (L) of a strip of material lengthens the charge carriers' path and therefore increases the total resistance of the device. Decreasing the area (A) through which charge carriers are channeled (outlined by width W and thickness T in the profile view of Fig. 2.1) has the same increasing effect on resistance R:

$$R = \frac{\rho L}{A} = \frac{\rho L}{WT} \tag{2.3}$$

where ρ is the resistivity of the material. Similarly, as temperature increases and atoms vibrate and electrons accelerate, collisions and therefore resistances increase, which is why most resistors have a positive temperature coefficient (i.e., resistance increases with temperature).

Since the material used to build a resistor intentionally (e.g., unsilicided polysilicon, lightly doped silicon, etc.) normally differs from that used to short-circuit the resistor to another device (e.g., silicided polysilicon, aluminum, etc.), the ends of the resistor strip include low-resistance interface contact areas (Fig. 2.1). Here, the effective width is increased and length decreased, relative to the resistor strip, to decrease

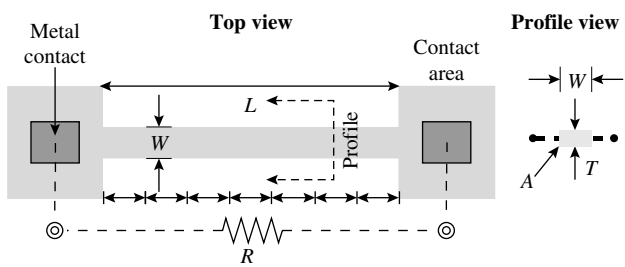


Figure 2.1 Resistor strip.

the overall resistance of the region and accommodate a metallic contact, resulting in the classical "doggy bone" structure outlined in the figure. A process engineer optimally defines the thickness of the material by, for instance, adjusting the deposition rate of polysilicon or the annealing time of diffusing dopant atoms, neither of which are under the control of an IC designer. Consequently, an IC designer normally specifies the top-view length L and width W of a resistor, not its thickness T, which is why IC designers think of *sheet resistance* R_S, not resistivity, when referring to the intrinsic resistance of a material,

$$R_S = \frac{\rho}{T} \tag{2.4}$$

Sheet resistance is convenient because it represents the resistance of one square (W equals L) of material, irrespective of its resulting area, which is why its symbol also takes the form of $R_{\square}$. For example, referring to Fig. 2.1, the total resistance, when neglecting the resistance of the contact areas, is approximately the resistance of seven squares in series, that is, 7 R_S:

$$R = NR_S \tag{2.5}$$

where N is the total number of squares in a strip of material.

2.1.2 Capacitors

The intrinsic advantage of integrated-circuit technologies is their ability to build numerous electronic devices (e.g., resistors) on a common, relatively low-impedance silicon substrate (e.g., p-doped silicon). For electrical isolation, a dielectric medium is inserted or created between these devices and their common substrate, the latter of which is connected to a low-impedance supply for the purpose of reverse-biasing all junctions in the IC. The end result, as it pertains to the resistor or link, is a parallel-plate capacitor between the entire device and substrate, as shown in Fig. 2.2. Its total capacitance

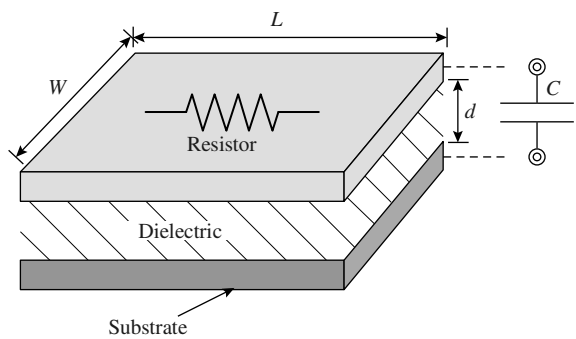


Figure 2.2 Capacitor to substrate.

increases with increasing surface area (i.e., increasing W and increasing L) and decreasing plate-separation distance d:

$$C = \left(\frac{W \cdot L}{d}\right)\varepsilon = \left(\frac{W \cdot L}{d}\right)k\varepsilon_o = W \cdot L \cdot C''_\varepsilon \tag{2.6}$$

where C''_ε is the capacitance per unit area and ε the permittivity of the dielectric medium, which is dielectric constant k times greater than permittivity of vacuum ε_o.

As with the resistor, the process engineer, not the IC designer, optimally sets the separation distance, leaving again width W and length L as flexible design parameters. In the case of diffusion resistors, however, the dielectric is the depletion layer created by reverse biasing the resistor-substrate's pn junction, the separation of which is dependent on the bias voltage across the junction, as discussed later in the pn-junction diode subsection. Ultimately, in the context of resistors and short-circuit links, these capacitors are parasitic and their effects are unwelcome. On the other hand, replacing the high-resistance material with a more conductive medium like aluminum, silicided polysilicon, or highly doped silicon (e.g., n+ doped silicon); increasing the surface area; and conforming the device into one or only a few squares reduce the resistive component of the device and enhance its capacitive effects, in the end building what amounts to a monolithic *capacitor*.

From a circuit perspective, the impedance across the capacitor (z_c) decreases with decreasing plate-separation distance d, increasing permittivity (i.e., higher capacitance), and increasing frequency, and is considered infinitely high at dc,

$$z_C = \frac{dv_C}{di_C} = \frac{dt}{C} = \frac{1}{sC} \tag{2.7}$$

where dt represents time and s in Laplace transforms is the equivalent of frequency. Consequently, the higher frequency components of a propagating signal will have a tendency to shunt through the capacitor, losing some of its energy in the process. These effects are parasitic and unavoidable in a resistor. Similarly, the ohmic voltage drops and power losses associated with the resistive component of a capacitor are parasitic effects.

2.1.3 Layout

A good resistor is one whose resistance is easy to define and capacitance to substrate negligibly small. The resistance of the contact areas, whose value is relatively more difficult to predict, should also be negligibly small, that is, large and no more than one square's worth in length. The dielectric used (e.g., silicon dioxide) should be thick with

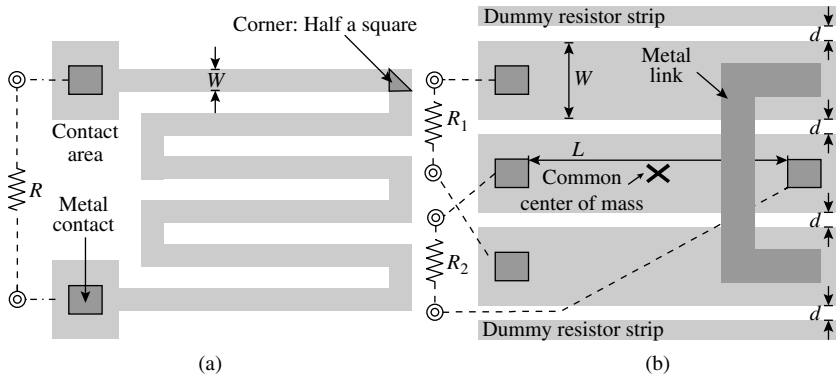


Figure 2.3 (*a*) High-value serpentine and (*b*) matched resistors (i.e., $R_1 \equiv 2R_2$).

a relatively high dielectric constant (e.g., field oxide). As such, a long, thin strip of high-resistivity material (e.g., 500 squares of 1 kΩ/□ of sheet resistance) over a thick dielectric constitutes a high-value resistor (e.g., 500 kΩ) with minimal parasitic capacitance. Such a high-value resistor is normally snaked into a *serpentine*, as shown in Fig. 2.3*a*, to contain its entire length within workable modular areas. Each corner of the serpentine, since charge-carriers traverse across the least resistive portions of the square, introduce approximately half a square worth of resistance.

The tolerance of a resistor mostly depends on the resistivity of the device (e.g., doping concentration) and normally exhibits a ±20% variation across process variations. Because of this relatively poor tolerance, IC designers rely on the matching performance of similarly built devices, the resistivities of which track better across an entire silicon die. Matching performance, since resistivities track, depends on the uniformity of the resistor widths across their entire lengths. Wider resistors (e.g., 3–5 times the minimum width allowed) normally match better (e.g., ±1%) because their respective width variations (e.g., 100 nm) constitute a smaller fraction of their total widths (e.g., 10 μm).

In the case of polysilicon resistors, the width variation depends on the uniformity of the etching rate across its length, which, as it turns out, depends on what is next to it. As such, designers normally ensure equidistant, adjacent resistor strips of the same material surround all resistor strips, as shown in Fig. 2.3*b*, where the outer resistor strips are thin dummy devices just for the sake of minimizing etching errors. However, it is impossible to present an equidistant dummy resistor strip to the inner side of a corner, unless its separating distance d is sufficiently large, in which case it would be impossible to present a dummy resistor strip to the extreme ends of the inner side. As a result, corners are normally avoided and metallic connections

are used to link resistor segments, as shown in Fig. 2.3*b*, unless silicon real-estate demands supersede matching performance concerns.

As a side note, to improve the matching performance of two resistors (e.g., R_1 and R_2), their constituent resistor segments can be *interdigitated*, as also shown in Fig. 2.3*b*, so as to average the effects of process-induced gradients across the die (e.g., doping concentration) on the resistors. If possible, for similar reasons, the resistors should also have a common center of mass and conform to a *common-centroid* configuration, as also illustrated in the figure. Finally, increasing the number of resistor segments and inter-digitating them increases the statistical granularity and resolution of the layout, further improving matching performance.

As with the resistor, a good capacitor is relatively easy to define with a resistive component that is negligibly small. Surface areas should therefore be large enough to ensure the resistances across the plates are low and the orthogonal parallel-plate capacitance dominates over fringing edge effects. The width of the dielectric material should be relatively shallow (e.g., thin silicon dioxide or thin oxide for short) and its dielectric constant high. As such, the surface areas of capacitors are normally high and conform to square-like shapes, as shown in Fig. 2.2. Their tolerance depends on the homogeneity of the dielectric constant from wafer to wafer and lot to lot, which is normally on the order of ±20%. As with resistors, designers prefer to rely on their matching performance, if possible, in other words, on the uniformity of edge variations (and consequential area variations) with respect to the overall areas. Consequently, large-area capacitors and, because of proximity etching errors, capacitors with equidistant dummy peripheral strips match better, as illustrated in Fig. 2.4.

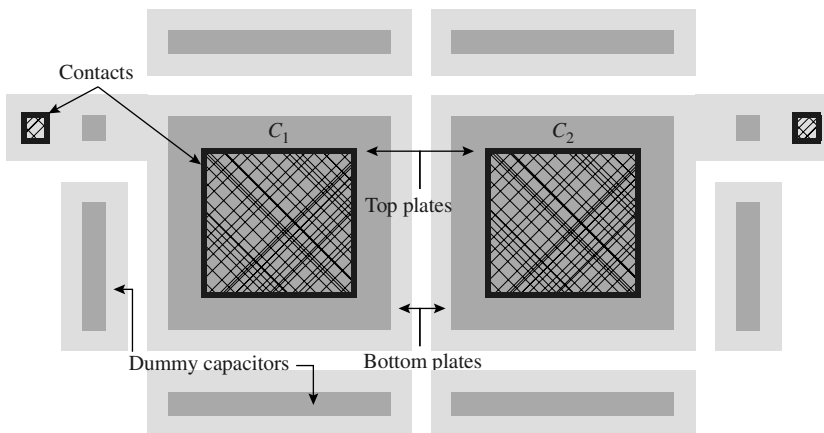


Figure 2.4 Matched capacitors (i.e., $C_1 \equiv C_2$).

As stated earlier, capacitance depends on the overlapping surface area between the top and bottom plates of a parallel-plate capacitor so etching errors on the top plate, as the plate that defines surface area A, are more important than errors on the bottom plate. To that end, ensuring proximity effects remain uniform across a matching capacitor array amounts to adding dummy top-plate strips to the surrounding periphery. Using dummy capacitors in place of dummy strips is better, however, as shown in Fig. 2.4, because the vertical placement of a plain poly-2 strip does not necessarily correspond with the top plate of a poly-poly capacitor. To be more explicit, field oxide separates a plain strip of poly-2 material from substrate whereas thin oxide separates the poly-2 plate from its poly-1 counterpart and field oxide separates the latter from substrate. These dummy devices may remain open-circuited but connecting them to low-impedance nodes is often preferred to shunt incoming noise away from the matching capacitor array.

2.2 PN-Junction Diodes

2.2.1 Large-Signal Operation

Intrinsic Device

The most appealing aspect of a pn-junction diode, although not always necessarily used for this purpose, is its unidirectional current-conducting feature (i.e., current only flows in one direction). The device is built by stacking two oppositely doped (i.e., p and n) pieces of silicon, as shown in Fig. 2.5*a*. Since the concentration gradients of

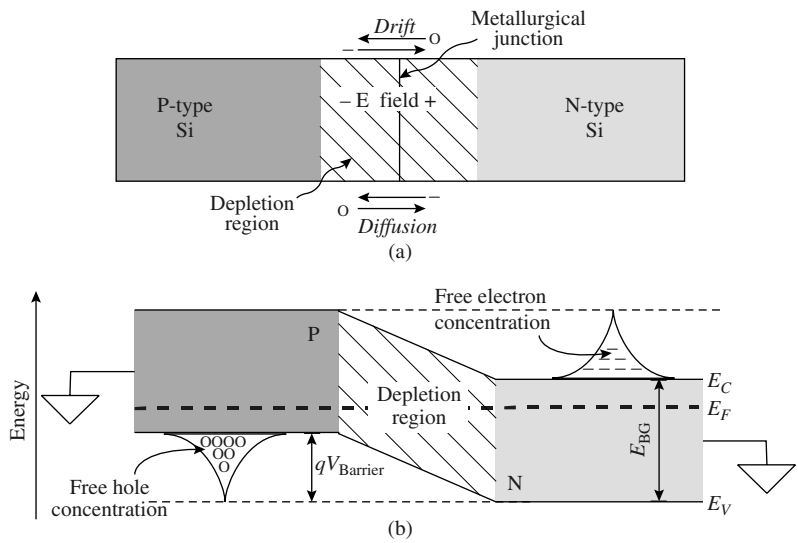


Figure 2.5 (*a*) Physical-profile view and (*b*) energy-band diagram of a pn-junction diode.

free electrons and holes (i.e., holes are counterparts to electron and therefore equivalent to "electron voids") across the junction are relatively high, electrons and holes diffuse from high- to low-concentration regions, that is, electrons migrate from n to p and holes from p to n regions. These migrating charge carriers leave their otherwise neutral home sites ionized (i.e., n region becomes positively charged and p negatively charged) and create in the process a depletion region (i.e., area depleted or void of charge carriers) in their respective home sites and an electric field whose induced current opposes that of diffusion. As a result, since the device is in equilibrium (i.e., there is no net current flow), diffusion and drift currents are equal and opposite.

The energy-band diagram shown in Fig. 2.5*b* is useful in describing the operation of the device because it conveys important information about its physical properties and electrical state. To start, the top line denotes the onset of the *conduction band* E_C, the energy level above which electrons are free to roam as *charge carriers*, and the bottom line the onset of the *valence band* E_V, the energy level below which electrons are strongly bound to the atom. Just as electrons above E_C are charge carriers, holes below E_V, given their complementary features, are also charge carriers.

The energy separation between the two bands is constant for a given material and commonly referred to as the *band gap* (E_{BG}). *Fermi level* E_F, which lies between E_C and E_V, is an abstract variable representing the energy level where a 50% probability of finding a charge carrier exists—this probability decreases exponentially for increasing and decreasing energy levels. This is not to say, however, a charge carrier exists at or around E_F, because none can reside in the band gap, only outside the bands. This does mean, however, the presence of charge carriers above E_C or below E_V decreases exponentially at energy levels that lie further away from E_F. Consequently, if E_F is closer to E_C, the probability of finding an electron charge carrier *above* E_C is greater than finding a hole below E_V, and it decreases exponentially at higher energy levels, as graphically illustrated in Fig. 2.5*b*. Similarly, a p-type material has E_F closer to E_V and its hole charge-carrier concentration peaks just below E_V and decreases exponentially at lower energy levels.

At equilibrium, E_F is flat across the pn junction and there is no net current flow through the diode. The conduction and valence bands through the mostly p- and n-type materials (shaded regions in the figure—bulk material) are therefore flat. However, as the material is depleted of charge carriers near the *metallurgical junction*, it effectively becomes less doped and E_F is consequently closer to midway across the band gap, behaving more like an *intrinsic* (i.e., undoped) semiconductor, the result of which is the band bending shown in Fig. 2.5*b*. The bending represents the potential energy barrier a charge carrier must overcome to diffuse across the junction, which is often times described in terms of *barrier potential* $V_{Barrier}$.

Applying a voltage across the diode shifts its barrier potential and forces it out of equilibrium. A positive voltage v_D on the p side, for instance, as shown in Fig. 2.6*a*, reduces the barrier and therefore exposes an *exponentially* increasing number of carriers over the barrier threshold, allowing them to diffuse to the oppositely doped side and in the process induce current flow ($i_D \propto \exp v_D$). With a reverse voltage v_R, on the other hand, as shown in Fig. 2.6*b*, the barrier increases, impeding diffusion and therefore current flow. The end result is a device that conducts, as shown in Fig. 2.6*c*, (1) an exponentially increasing current i_D into the p side (*anode*) of a pn junction with an increasing *forward-bias*

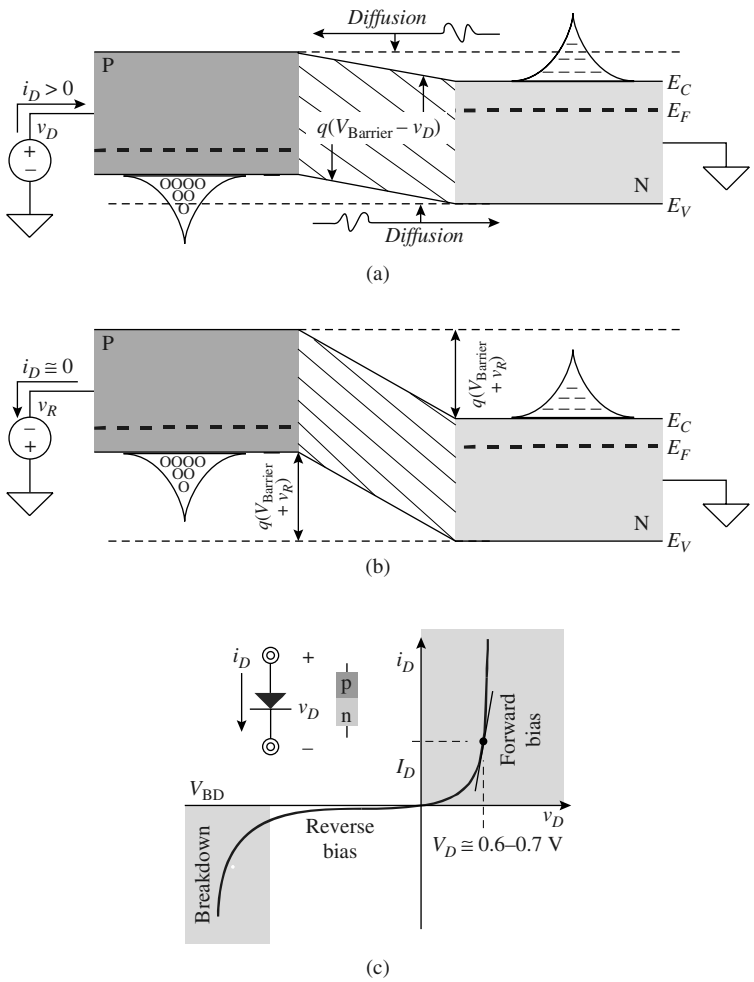


Figure 2.6 Energy-band diagrams of a pn-junction diode during (*a*) forward- and (*b*) reverse-bias conditions and (*c*) the resulting current-voltage (I-V) curve.

voltage v_D, (2) zero current with zero voltage, and (3) little to no current with a *reverse-bias* voltage (v_D is below 0 V):

$$i_D = I_S\left[\exp\left(\frac{v_D}{V_t}\right) - 1\right] \tag{2.8}$$

where I_S is the *reverse-saturation current* constant for the material and V_t the *thermal voltage*, the former of which is on the order of femto-amps (10^{-15}) and the latter approximately 25.6 mV at room temperature. As a side note, high temperatures aid the diffusion process (via *diffusion coefficient* and *mobility*) and therefore induce higher current levels (i.e., I_S increases), which is equivalent to saying, for a given current density, diode voltage v_D decreases with increasing temperatures—the diode is "stronger" at higher temperatures and v_D has a negative temperature coefficient.

If a sufficiently large reverse voltage exists across the junction, the device breaks down and conducts current in the normally off direction (i.e., i_D reverses direction, see Fig. 2.6*c*). Under these conditions, past *breakdown voltage* V_{BD}, one or a combination of two breakdown mechanisms results that induces current flow. In a highly doped pn junction, for instance, where the region near the metallurgical junction is relatively difficult to deplete of carriers because of sheer abundance, the depletion width is narrow and the resulting electric field across it high, inducing electrons to *tunnel* through and across the region, as illustrated in Fig. 2.7. In lightly doped junctions, the depletion width is

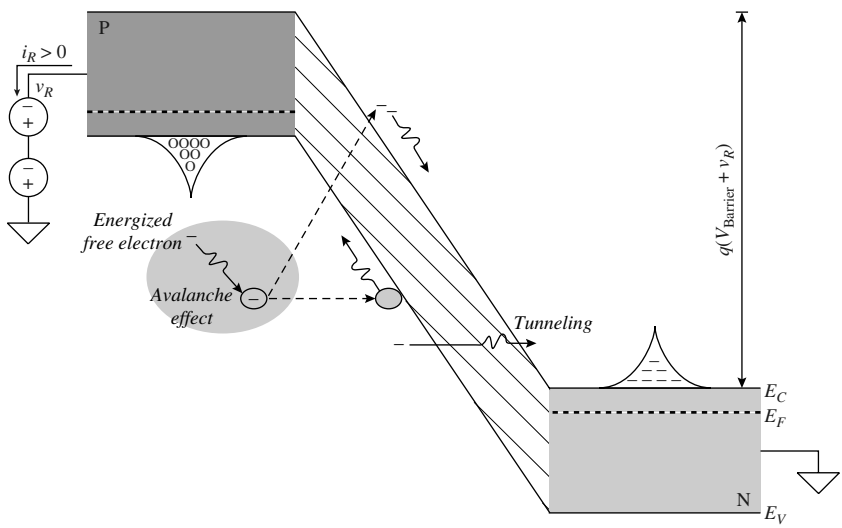


Figure 2.7 PN-junction diode in its breakdown region.

longer and therefore impede the tunneling effect but increasing reverse-bias voltage v_R further can increase the electric field to the point where free electrons energize and accelerate into otherwise bounded electron-hole pairs in the valence band with sufficient energy to break them apart through a process called *impact ionization*. The number of free charge carriers consequently increase (Fig. 2.7), inducing current flow. Since one free electron frees another (and a hole), the process multiplies the number of free carriers in the system in *avalanche* fashion, which translates to increasing current flow.

Most moderately doped pn-junction diodes exhibit a combination of both tunneling and avalanche effects with breakdown voltages of 6–8 V. Higher doped devices, which have narrower depletion widths, have lower breakdown voltages and yield more tunneling than avalanche current, and vice versa. Irrespective of the mechanism, however, diodes optimized for this region of operation are known as *zener* diodes and their respective zener voltages equate to their breakdown voltages (V_{BD}'s). As a side note, subjecting a diode to extreme and sustained high-voltage conditions in either reverse ($V_R >> V_{BD}$) or forward ($V_D >>$ 0.6–0.7 V) bias induces so much current the material and the package in which it is housed strain beyond their power and thermal limits, exposing it beyond its safe-operating area (SOA) and irreversibly damaging it.

Upon further scrutiny, as in the case of a parallel-plate capacitor, the bulk p- and n-type materials of a reverse-bias pn-junction diode comprise two conductive parallel plates separated by a dielectric medium (i.e., depletion region). The depletion capacitance ($C_{j.dpl}$) that results, as in parallel-plate capacitors, increases with increasing cross-sectional areas (A's) and decreasing depletion widths. Highly doped regions and lower reverse-bias voltages v_R's (or higher forward-bias voltages v_D's) produce shorter depletion widths because dense regions are more difficult to deplete of carriers and lower v_R's attract less carriers away from the metallurgical junction to the bulk material

$$C_{j.\mathrm{dpl}} = \frac{AC''_{\mathrm{jo}}}{\sqrt{1 - \dfrac{v_D}{V_{\mathrm{BI}}}}} \tag{2.9}$$

where V_{BI} is the built-in barrier potential (roughly 0.6 V) and C''_{jo} the zero-bias (i.e., v_D is zero) capacitance per unit area, which is a process and junction-dependent parameter—C''_{jo} depends on the doping densities of both materials, but more so on the less doped side because it normally dominates the overall width. Note the peripheral sidewall of the device contributes additional (and similar) capacitance.

Even in forward-bias conditions, small as it may be, the depletion region remains and depletion capacitance $C_{j.dpl}$ continues to increase. However, since voltage-dependent charge carriers now diffuse across

the junction in a finite forward-transit time τ_F (that is process dependent), diffusion component $C_{j.\text{diff}}$ increases the overall capacitance of the junction by

$$C_{j.\text{diff}} = \frac{\Delta q}{\Delta v_D} = \left(\frac{di_D}{dv_D}\right)dt = g_d \tau_F \tag{2.10}$$

where g_d is the effective conductance of the diode (i.e., first derivative of diode current i_D with respect to its voltage v_D) and total junction capacitance C_j is

$$C_j = C_{j.\text{diff}} + C_{j.\text{dpl}} \tag{2.11}$$

Since the conductance of the pn-junction diode is considerably high with forward-bias voltages, the diffusion component normally overwhelms depletion capacitance $C_{j.\text{dpl}}$ in forward-bias conditions, and vice versa. Ultimately, however, irrespective of the region of operation, junction capacitance C_j exists across intrinsic diode D, as depicted in the complete large-signal model shown in Fig. 2.8*a*.

Extrinsic Device

Maneuvering technology to build a diode necessarily introduces parasitic junctions and components to the intrinsic device. To start, the diode is a series combination of oppositely doped silicon pieces each of which introduces series parasitic resistances (R_P and R_N in Fig. 2.8*a*). More importantly, however, all devices in an integrated circuit (IC) share a common p- or n-type substrate, isolation from which is normally achieved by reverse biasing all substrate pn junctions, resulting in parasitic depletion capacitances C_{jN} and C_{jP} in p- and n-type substrates, respectively. For instance, diffusing p-type material into an already diffused n-type well to build an intrinsic pn-junction diode (Fig. 2.8*b*) inherently exposes the n *cathode* to the p-substrate. The current of the resulting extrinsic pn-junction diode must be negligibly

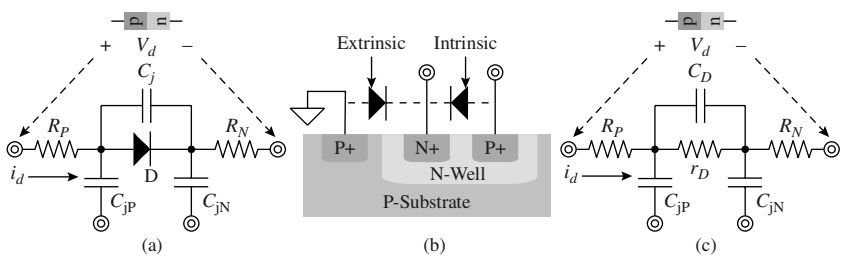


Figure 2.8 Complete (*a*) large-signal model, (*b*) physical profile view, and (*c*) small-signal model of a pn-junction diode.

small for proper intrinsic diode operation, which is why p-type substrates are connected to negative supply rails and n-type to positive supplies.

2.2.2 Small-Signal Response

AC diode voltage and current signals v_d and i_d refer to small enough variations in diode voltage v_D and current i_D to constitute only a considerably small fraction of their dc biasing counterparts V_D and I_D. Given the low-magnitude nature of an ac voltage, a linear approximation of the exponential diode-current relation (i.e., slope or first derivative at biasing point I_D-V_D in Fig. 2.6*c*) correlates reasonably well with the actual exponential response. As a result, ac current i_d, with the linearized model discussed, is linearly proportional to ac diode voltage v_d:

$$i_d = v_d\left(\frac{di_D}{dv_D}\right) = v_d\left(\frac{I_S}{V_t}\exp\frac{v_D}{V_t}\right) \approx v_d\left(\frac{I_D}{V_t}\right) \equiv v_d g_d \equiv \frac{v_d}{r_d} \qquad (2.12)$$

where g_d is the effective ac conductance and r_d the equivalent ac resistance of the diode. The small-signal response of a diode can therefore be described by a single resistor whose resistance is the ratio of thermal voltage V_t and dc biasing current I_D (i.e., $1/g_d$). As with large signals, the bulk resistances associated with the p- and n-type silicon strips introduce additional parasitic resistances R_P and R_N to the device, as included in the complete small-signal model shown in Fig. 2.8*c*. Normally, the effects of the extrinsic components (i.e., R_P, R_N, C_{JP}, and C_{JN}) are negligibly small and therefore neglected for first-order analysis and design, but not for simulations and worst-case corner analysis.

2.2.3 Layout

As with resistors and capacitors, proximity, orientation, common-centroid, inter-digitaion, and cross-coupling techniques improve the matching performance of diodes. Unlike polysilicon resistors and capacitors, however, they do not suffer from etching errors, as they are diffused devices. Consequently, adding dummy devices (or diffusion strips) around the periphery of an array of diodes may not improve matching performance to the same extent dummy polysilicon strips around polysilicon devices do. However, out diffusion is affected by the doping concentration around its surrounding regions and adding dummy, equidistant diffusion strips therefore improves the uniformity of the out-diffused areas and the parasitic components attached to them. Figure 2.9 illustrates the top layout views of readily available pn-junction diodes in CMOS and bipolar or biCMOS technologies: p+-source-drain (p+-S/D) n-well and n+-emitter p-base diodes.

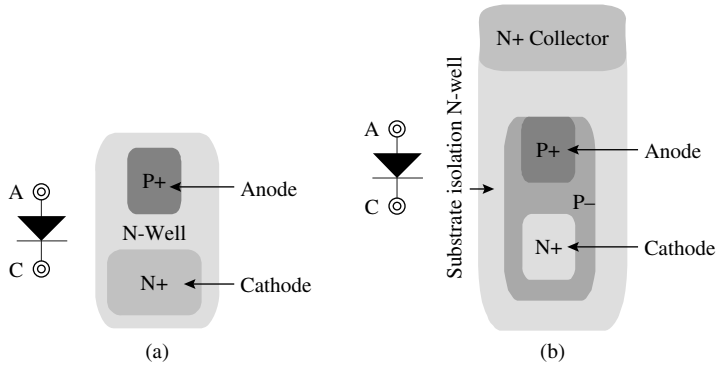


Figure 2.9 Top layout views of (*a*) CMOS p+-source-drain (p+-S/D) n-well and (*b*) bipolar or biCMOS n+-emitter p-base pn-junction diodes.

2.3 Bipolar-Junction Transistors

2.3.1 Large-Signal Operation

Intrinsic Device

A *bipolar-junction transistor* (BJT) comprises two back-to-back diodes, as inferred from the energy-band diagram shown in Fig. 2.10, with a short common middle region called the *base*. The base derives its name from the first physical implementation of the device, because it served as the physical base of the prototyped transistor. As in diodes,

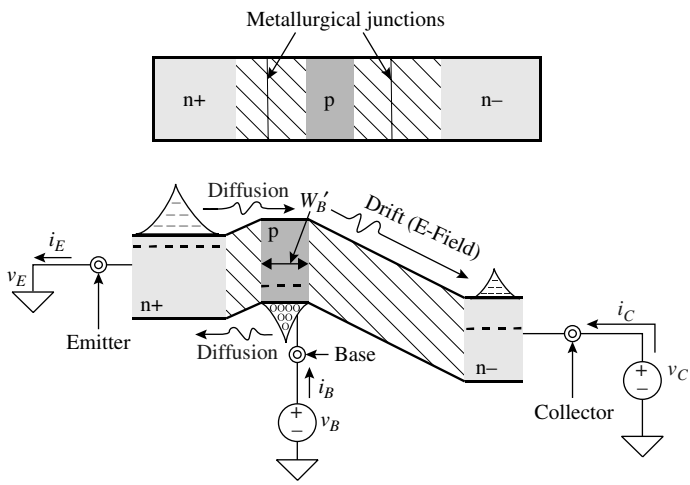


Figure 2.10 Energy-band diagram of an NPN BJT in the forward-active region.

charge carriers diffuse across the base-emitter junction when forward biased (i.e., $v_B > v_E$). Unlike diodes, however, because the base width (W_B') is relatively short when compared to the average diffusion length, the electric field across the reverse-biased base-collector junction (i.e., $v_C > v_B$) sweeps the minority charge carriers across the base to the collector before they have a chance to recombine in the base. As a result, as graphically shown in Fig. 2.10, most of the *emitted* electrons from the n+ emitter region are *collected* by the n-type terminal as collector current i_C. The holes that diffuse from the base to the emitter as a result of the forward-biased base-emitter junction produce a base current i_B, yielding a total emitter current that is equal to the aggregate sum of currents i_C and i_B.

Ideally, as the name implies, the transistor is a resistor with a conductance-modulating third terminal, the base. The input current into the base should therefore be low and the collector-base current ratio high, which is why the emitter is normally more doped (e.g., n+ or p+) than the base (e.g., p or n). Less doping in the collector (e.g., n− or p−) discourages reverse-active currents and increases the base-collector breakdown voltage, allowing the device to sustain higher collector-emitter voltages. Since the current flowing through the collector is the diffused minority charge-carrier current into the base, it is exponential with respect to forward-bias base-emitter voltage v_{BE}, as shown in the I-V curves of Fig. 2.11*b*. Because the base-collector depletion region increases with increasing reverse-bias collector-base voltage v_{CB}, base width W_B' decreases with increasing collector-emitter voltage v_{CE}, and decreasing the base width decreases the number of recombined minority carriers in the base and therefore slightly increases collector current i_C:

$$i_C = \left(I_S \exp \frac{v_{BE}}{V_t}\right)\left(1 + \frac{v_{CE}}{V_A}\right) \tag{2.13}$$

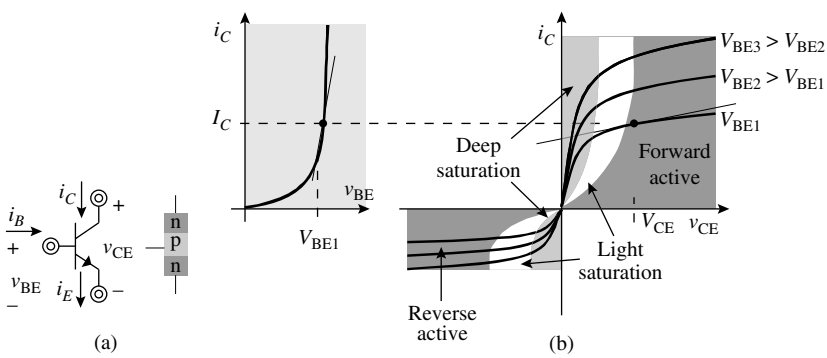


FIGURE 2.11 (*a*) Symbol and (*b*) I-V curves of an NPN BJT.

where V_A refers to the *base-width modulation* effect, which is otherwise known as *Early voltage*. The collector-base current ratio is the current gain (β) of the device,

$$i_C = \beta i_B \tag{2.14}$$

Note the arrow corresponds to the optimized emitter and points in the direction of current flow across the base-emitter junction.

The BJT has four regions of operation. From an analog perspective, the most appealing mode is the one just described, when forward biasing the base-emitter junction and reverse biasing the base-collector junction, because it yields the highest current- and voltage-gain characteristics. Process design engineers therefore optimize these devices to operate in this specific region, in the *forward-active* mode. Inversing the biasing conditions of the emitter and collector produces similar current-flow dynamics as in forward active but with less attractive results. As in forward active, forward biasing the base-collector junction induces carrier diffusion from collector to base and reverse biasing the base-emitter junction sweeps the diffused minority carriers into the emitter. Unlike forward active, however, the magnitude of the currents and their associated gain are lower because the collector has fewer free charge carriers when compared to the emitter; in other words, it has lower doping concentration.

Forward biasing the base-collector junction while also forward biasing the base-emitter junction induces base-collector current flow, decreasing (and *saturating*) the effective current gain of the BJT when compared to the forward-active gain. In practice, however, slightly forward biasing the base-collector junction (i.e., *slightly saturating* the transistor) by, for example, 0–0.3 V induces negligible additional base current and therefore retains the high current-gain characteristics of the forward-active region. Reverse biasing both junctions, however, results in negligible current flow in all terminals and places the device in its *off* mode. It is important to note, as in the diode, that higher temperatures aid the diffusion process (i.e., increases reverse-saturation current I_S by increasing minority-carrier diffusion coefficient and mobility), allowing lower base-emitter voltages to induce higher collector currents.

Extrinsic Device

The intrinsic transistor is literally the two back-to-back pn junctions combined, not the diffusions and contacts used to connect the device (Fig. 2.12). The series resistances associated with the ohmic contacts of each terminal and their respective parasitic pn junctions comprise the extrinsic portion of the BJT. As such, series base resistance R_B, emitter resistance R_E, and collector resistance R_C are present in both large- and small-signal models. Similarly, the reverse-biased collector-substrate pn-junction diode and its equivalent capacitance (C_S) are also present (Fig. 2.13*a*).

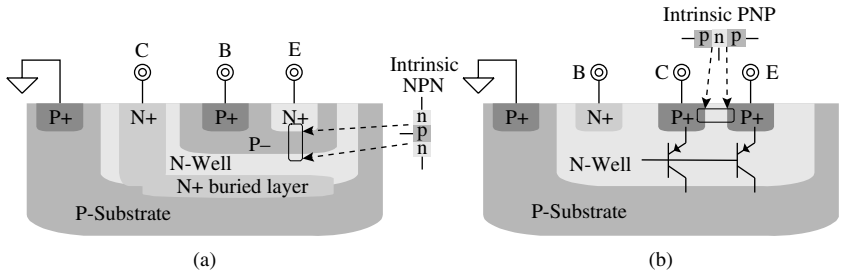


Figure 2.12 Physical profile view of typical (*a*) vertical NPNs in standard bipolar and biCMOS and (*b*) lateral PNPs in standard CMOS process technologies.

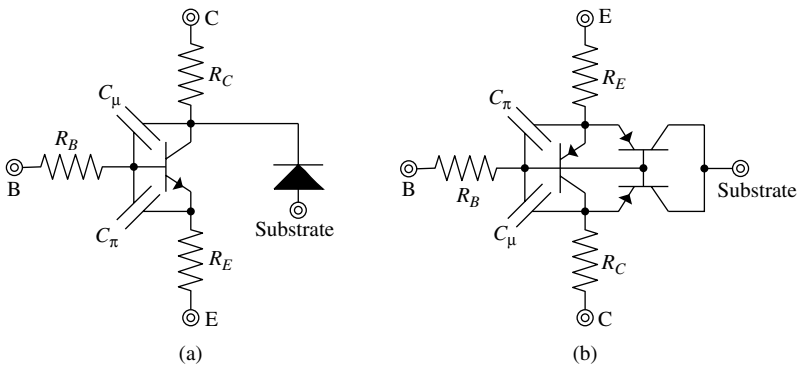


Figure 2.13 Large-signal models of (*a*) vertical NPN and (*b*) n-well lateral PNP transistors.

The n-well lateral PNP also has two parasitic bipolar PNP transistors sharing the same base as the lateral device and having a substrate tied to ground through the substrate (Figs. 2.12*b* and 2.13*b*). Since the emitter is at a higher potential, its respective parasitic transistor is more problematic, decreasing the overall current efficiency of the lateral device. Adding an n+ buried layer to the n-well, if available, increases recombination and therefore decreases the gain of these parasitic devices.

The buried layer shown in the standard bipolar or biCMOS vertical NPN in Fig. 2.12*a* decreases the parasitic series resistance associated with the collector. Without this layer, since the intrinsic collector is deep within the substrate and relatively far away from its extrinsic surface contact, the device enters the saturation region at a relatively higher extrinsic collector-emitter voltage v_{CE} because v_{CE}'s intrinsic counterpart is lower by an R_C-ohmic drop ($v_{C'E} = v_{CE} - i_C R_C$). Emitter resistance R_E is normally low because it is close to its extrinsic contact

point, to the surface of the die. The standard CMOS lateral PNP shown in Fig. 2.12*b* normally has lower collector resistance because of the proximity of its intrinsic collector to its extrinsic contact. Unfortunately, however, it suffers from larger base resistance and higher base-width modulation effects (i.e., lower Early voltage V_A or equivalently lower output resistance r_o) because its base has lower doping concentration than in an optimized vertical device, and lower doping concentration produces higher depletion-width variations when confronted with changes in reverse-bias voltages.

2.3.2 Small-Signal Response

For the most part, analog circuit designers bias the transistor at a particular dc quiescent point (e.g., V_{BE1}, V_{CE}, I_B, and I_C in Fig. 2.11) and mix and sample small ac signal variations (e.g., v_{be}, v_{ce}, i_b, and i_c) around that point. Because these variations are small, linear first-order model approximations predict their behavior reasonably well. As such, a small change in v_{BE} produces roughly linear variations in i_C and i_B and changes in v_{CE} additional variations in i_C, as illustrated in Fig. 2.14*a*.

The product of the first partial derivative of i_c with respect to v_{be} ($\partial i_C/\partial v_{BE}$) and small variation v_{be} describes the resulting change in i_c:

$$i_c\Big|_{v_{ce}=0} \approx v_{be}\left(\frac{\partial i_C}{\partial v_{BE}}\right) \equiv v_{be}g_m \approx v_{be}\left(\frac{I_C}{V_t}\right) \tag{2.15}$$

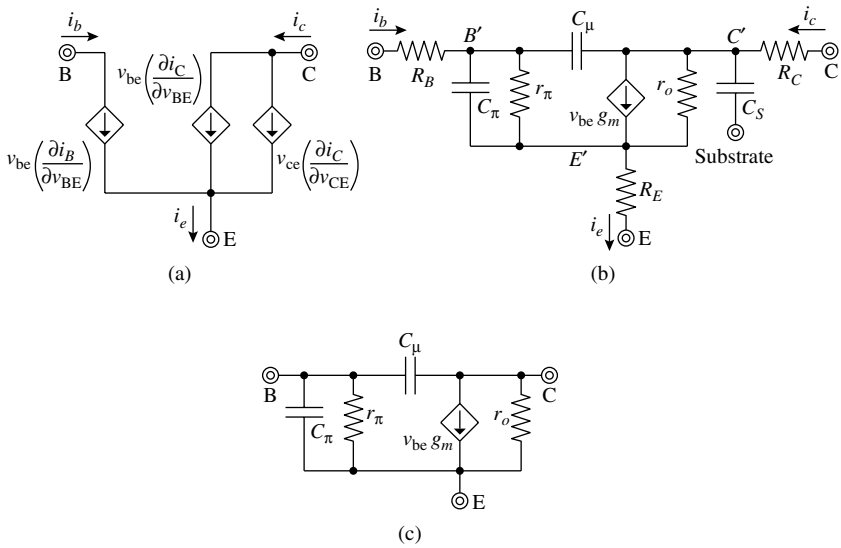


Figure 2.14 (*a*) Small-signal current variations model and (*b*) complete and (*c*) simplified small-signal models of an NPN BJT.

where g_m, as also shown in Fig. 2.14*b*, is the effective transconductance and I_C the biasing dc current of the BJT. Similarly, the product of the first partial derivative of i_c with respect to v_{ce} ($\partial i_C/\partial v_{CE}$) and small variation v_{ce} describes the resulting change in i_c:

$$i_c\Big|_{v_{be}=0} \approx v_{ce}\left(\frac{\partial i_C}{\partial v_{CE}}\right) \equiv v_{ce}g_o \approx v_{ce}\left(\frac{I_C}{V_A}\right) \equiv \frac{v_{ce}}{r_o} \tag{2.16}$$

where g_o is the effective output conductance of the device. Because this linear current variation results from a change in the voltage across its own two terminals v_{ce}, output resistor r_o, as shown in Fig. 2.14*b*, with an equivalent resistance of V_A/I_C can be used in place of dependent conductance g_o. Finally, the product of the first partial derivative of i_b with respect to v_{be} ($\partial i_B/\partial v_{BE}$) and small variation v_{be} describes the resulting change in i_b, which is also the resulting change in i_c divided by the current gain ratio β:

$$i_b \approx v_{be}\left(\frac{\partial i_B}{\partial v_{BE}}\right) = \frac{v_{be}}{\beta}\left(\frac{\partial i_C}{\partial v_{BE}}\right) = v_{be}\frac{g_m}{\beta} \equiv \frac{v_{be}}{r_\pi} \tag{2.17}$$

Again, since the variation in i_b is the result of a change across its own two terminals, resistor r_π, as shown in Fig. 2.14*b*, with equivalent resistance g_m/β also predicts i_b.

Because a BJT is two back-to-back diodes, base-emitter and base-collector diode capacitances are also present in the device. The base-emitter capacitance is a stronger function of diffusion than forward-biased depletion width and therefore linearly proportional to the time minority carriers take to cross the base (i.e., forward transit time τ_F) and transconductance g_m:

$$C_{BE} = C_{j.dif} + C_{j.dpl} \approx C_{j.dif} = \frac{\Delta q}{\Delta v_{BE}} = \left(\frac{\Delta i_C}{\Delta v_{BE}}\right)\tau_F = g_m\tau_F \equiv C_\pi \tag{2.18}$$

where q is charge and C_π the base-emitter junction capacitance. The base-collector capacitance, on the other hand, because the junction is reverse biased, is only a function of depletion width. As such, this capacitance is inversely proportional to reverse-bias collector-base voltage v_{CB} and linearly proportional to area A and zero-bias junction capacitance per unit area C''_{jcbo}:

$$C_{BC} = \frac{AC''_{jcbo}}{\sqrt{1+\dfrac{v_{CB}}{V_{BI}}}} \equiv C_\mu \tag{2.19}$$

where V_{BI} refers to the built-in barrier potential and C_μ to base-collector capacitance C_{BC}. The effects of extrinsic components R_B, R_E, R_C, and C_S are normally negligibly small, when considering first-order effects, which is why the simplified small-signal model of Fig. 2.8*c* is often useful. As in the diode, peripheral sidewall contributes additional and similar capacitance.

Base-emitter and base-collector capacitances C_π and C_μ shunt base input signal v_{be} at higher frequencies and consequently decrease the effective transconductance gain of the device. The frequency when short-circuit current gain i_c/i_{in} falls to one is defined as the *transitional frequency* (f_T) of the device beyond which the transistor is, for the most part, of no use, that is, incapable of amplifying ac signals. During short-circuit conditions, when small-signal v_{ce} is zero, r_π, C_π, and C_μ are in parallel (neglecting parasitic ohmic resistors R_B, R_C, and R_E in Fig. 2.14*b*) and the voltage across them (v_{be}) is the product of input current i_{in} and its equivalent parallel impedance,

$$\left.\frac{i_c}{i_{in}}\right|_{v_c=0} = \left.\frac{v_{be}g_m}{i_{in}}\right|_{v_c=0} = i_{in}\left[r_\pi \parallel \frac{1}{(C_\pi + C_\mu)s}\right]\left(\frac{g_m}{i_{in}}\right)$$

$$= \left[\frac{r_\pi}{1+r_\pi(C_\pi+C_\mu)s}\right] g_m \Big|_{f_T=\frac{g_m}{2\pi(C_\pi+C_\mu)}} \equiv 1 \qquad (2.20)$$

where i_c/i_{in} is evaluated at its transitional frequency f_T. Higher transconductances with smaller device areas (i.e., lower capacitance) yield higher transitional frequencies and are therefore deemed better (i.e., higher frequency) devices.

2.3.3 Layout

Figure 2.15 illustrates the top layout views of typical bipolar and biCMOS vertical NPN and CMOS lateral PNP transistors, both of which are direct translations of the profile views shown in Fig. 2.12. For good matching performance, multiple devices should be placed close to each other with a common center of mass and cross-coupled to cancel die-wide process gradients. Although not as important as in polysilicon resistors and capacitors, which suffer from etching errors, placing equidistant dummy devices around the periphery of an array of matched devices improves matching performance by emulating and mimicking the same peripheral diffusion conditions for all devices and therefore better matching their out-diffusion distances.

As for power, increasing the emitter area of a vertical BJT increases the current-carrying capabilities of the device. In the lateral case, however, the intrinsic emitter area is the peripheral sidewall of the emitter

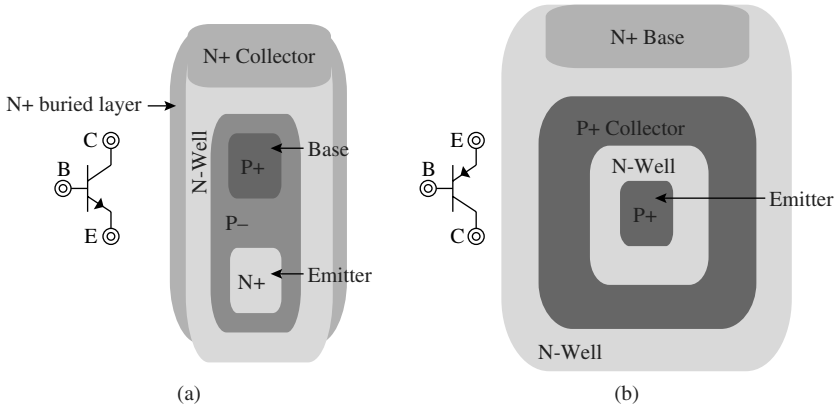


Figure 2.15 Top views of (*a*) vertical bipolar and biCMOS NPN and (*b*) lateral CMOS PNP BJTs.

diffusion, not the top-view area. As such, a "donut" (i.e., ringed) collector around a minimum-sized emitter dot collects more diffused minority carriers from the base (than a strip) and consequently improves the transport efficiency of the lateral device. To increase its current-carrying capabilities, unlike the vertical device, multiple, minimum-sized emitter-dot lateral devices are connected in parallel.

In analog applications, two or more transistors may share one or two of the three terminals available. It is possible, depending on the configuration and technology used, to merge two or more devices sharing two common terminals, optimizing the use of silicon real estate and reducing cost. A single vertical BJT, for instance, can have several emitter dots to channel and split its collector current into several emitter currents, as shown in the multiple-emitter vertical NPN device of Fig. 2.16*a*. More common, however, is the use of multiple

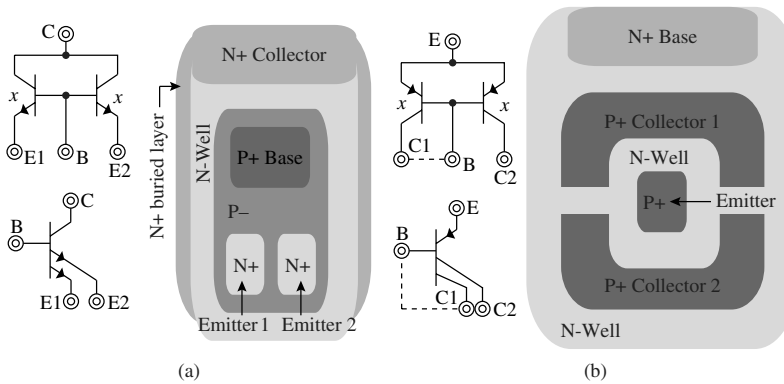


Figure 2.16 Top layout views of (*a*) multiple-emitter vertical NPN and (*b*) multiple-collector lateral PNP transistors.

collector devices in current-mirror configurations, which the merged, multiple-collector lateral PNP transistor shown in Fig. 2.16*b* achieves readily. The basic idea is to collect a fraction of the diffused carriers with one collector and the remainder with the other collector. Proportionally sizing the emitter areas in the vertical device and collector-ring ratio in the lateral case sets the current split ratio between emitter and collector terminals, respectively.

2.4 Metal-Oxide Semiconductor Field-Effect Transistors

2.4.1 Large-Signal Operation

Intrinsic Device

Metal-oxide semiconductor field-effect transistors (MOSFETs) differ from BJTs in that current flow is mostly the result of an electric field, not diffusion. They are two-terminal resistors with a conductance-controlling third terminal called the *gate*. The voltage applied to the parallel-plate metal-oxide-semiconductor (MOS) capacitor (gate) directly above the resistor modulates the charge (i.e., resistance) across the channel between the two terminals of the resistor. Even though modern-day MOSFETs use polysilicon material in place of the metal gate (Fig. 2.17), they conventionally retain the name MOSFET.

Applying a negative voltage to the gate ($v_G < 0$) of an n-channel MOSFET (i.e., NMOS transistor), as shown in Fig. 2.17*a*, attracts p+ majority carriers from the bulk to the surface near the gate, effectively *accumulating* charge carriers near the surface and increasing its

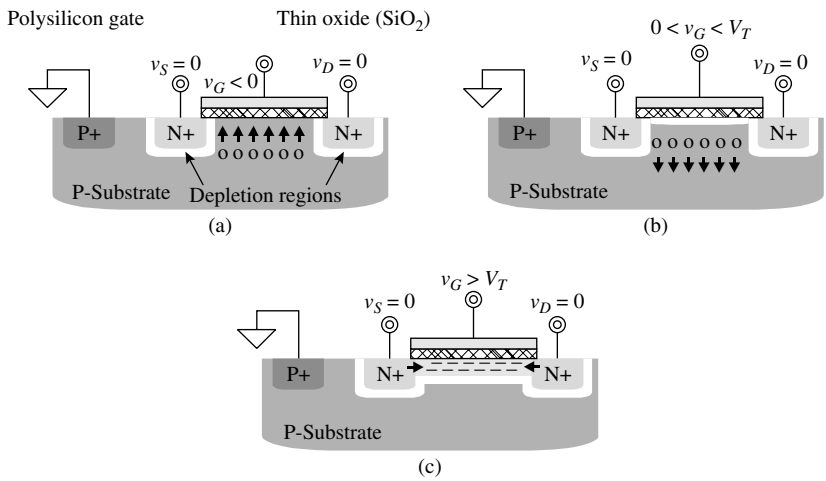


Figure 2.17 N-channel metal-oxide semiconductor field-effect transistor (MOSFET) in (*a*) accumulation, (*b*) depletion, and (*c*) inversion.

majority carrier concentration. Because all pn junctions are reverse biased, no current flows and the device is said to be *off*. Applying a small positive voltage to the gate, as shown in Fig. 2.17*b*, repels p+ majority carriers from the surface and *depletes* the region immediately beneath the surface of charge carriers, but still no current flow is established. Applying sufficient positive voltage to the gate, above the *threshold voltage* of the device (e.g., $v_G > V_T$), attracts charge carriers from the n+ contacts into the region just beneath the gate, effectively *inverting* the latter and connecting the n+ terminals with an n-type *channel* (Fig. 2.17*c*), which is why this device is called an "n-channel" MOSFET. Further increasing gate voltage v_G, increases the charge-carrier concentration of the channel and therefore increases the effective conductance of the same, achieving the desired operational dynamics of a transistor.

As with any resistor, the channel resistance is a function of minority carrier mobility and increases with increasing channel length L and decreasing channel width W. As previously stated, it also decreases with increasing gate voltage v_G (or gate-source voltage v_{GS}) and gate-oxide capacitance. Ultimately, the current through the device is inversely proportional to channel resistance R_{Channel} (and the dependencies just described) and directly proportional to the voltage across its terminals (i.e., drain-source voltage v_{DS}):

$$i_D = \frac{v_{DS}}{R_{\text{Channel}}} = v_{DS}\left(\frac{W}{L}\right)K_N'[(v_{GS} - V_T) - 0.5v_{DS}]\Big|_{v_{DS}<V_{DS(sat)}}$$

$$\approx \left(\frac{W}{L}\right)K_N'[(v_{GS} - V_T)v_{DS}] \tag{2.21}$$

where K_N' is the transconductance parameter, which is equal to the product of n-type mobility μ_N and oxide capacitance per unit area C_{OX}''. This relationship only holds true if the voltage across the channel (v_{DS}) is relatively low (i.e., below saturation voltage $V_{DS(sat)}$) and is said to describe the *triode* or *nonsaturated* region of the MOSFET (Fig. 2.18*a* and *b*).

The charge immediately beneath the gate is directly proportional to the voltage applied across the gate-oxide-semiconductor parallel-plate capacitor (i.e., $Q = C_{OX}V_{OX}$). As such, increasing the channel voltage v_{DS} while keeping gate-source voltage v_{GS} unchanged decreases the charge at the drain terminal, but not the source terminal, as graphically shown in Fig. 2.18*b*. When the voltage across the oxide capacitor at the drain is effectively zero, that is, when v_{GD} equals V_T or

$$v_{DS}\Big|_{\text{Pinch}} = v_{GS} - V_T \equiv V_{DS(sat)} \tag{2.22}$$

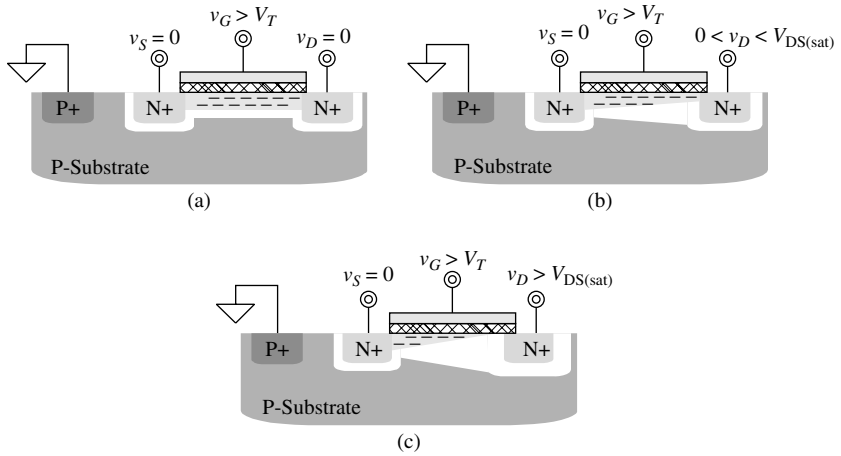


FIGURE 2.18 N-channel MOSFET in (*a*) and (*b*) triode and (*c*) saturation.

where $V_{DS(sat)}$ is the *saturation voltage* of the MOSFET, the channel *pinches* at the drain and additional increases in v_{DS} simply shift the pinch-off point further into the channel, as shown in Fig. 2.18*c*. The voltage across the now shorter channel remains constant at $V_{DS(sat)}$, which is why drain current i_D saturates to a relatively constant value, but because the channel length decreases slightly with increasing v_{DS}, channel resistance $R_{Channel}$ also decreases and i_D increases slightly with increasing v_{DS}:

$$\begin{aligned} i_D\Big|_{v_{DS}\geq V_{DS(sat)}} &= \frac{V_{DS(sat)}}{R_{Channel}} \\ &\approx V_{DS(sat)}\left(\frac{W}{L}\right)K_N'[(v_{GS}-V_T)-0.5V_{DS(sat)}](1+\lambda v_{DS}) \\ &\approx 0.5\left(\frac{W}{L}\right)K_N'(v_{GS}-V_T)^2(1+\lambda v_{DS}) \end{aligned} \tag{2.23}$$

where λ is the channel-length modulation parameter, the device is said to be in *saturation*, and $V_{DS(sat)}$ is $v_{GS} - V_T$ and can be rewritten as

$$V_{DS(sat)} \equiv V_{GS} - V_T \approx \sqrt{\frac{2I_D}{\left(\frac{W}{L}\right)K'}} \tag{2.24}$$

Since the current is relatively constant against variations in v_{DS}, except for small channel-length modulation effects, the effective output resistance of the transistor is high, which is helpful in high-gain circuits.

Just as the emitter and collector derive their names because they emit and collect minority charge carriers into and out of the base, respectively, source and drain derive their names because they *source* and *drain* minority carriers into and out of the channel. The source in an n-channel MOSFET is always the lower potential terminal and the drain the higher potential, and vice versa for the p-channel transistor, given the complementary nature of the device. Unlike the BJT, which device engineers optimize to operate in forward active, the MOSFET enjoys the benefit of true asymmetric performance because the source and drain are equally doped and can therefore be swapped without any degradation in performance, as shown in the I-V curves of Fig. 2.19. In saturation, the device follows a *square-law* behavior, given its square dependence to v_{GS}. Note there is no input current into the gate because the input impedance is purely capacitive. In addition, the arrow always corresponds to the source and its direction denotes the direction of current flow.

In an n-well p-substrate technology, the *bulk* terminal (also known as the *body*) of the n-channel transistor is the substrate, as shown in Figs. 2.17 and 2.18, so its potential is always the most negative available in the IC to reverse bias all substrate pn junctions. The bulk (or body) terminal of the p-channel transistor, on the other hand, is the n-well (Fig. 2.20), which can be connected to any potential the design engineer chooses. Normally, bulk and source are short-circuited (or bulk tied to the most positive potential) to prevent source-bulk and drain-bulk junctions from forward biasing.

Changing the bulk-terminal voltage modulates the conductance of the channel, which is why the bulk is also a *back gate*. The resulting drain-current variation is called the *bulk effect*. In the case of a p-channel MOSFET (Fig. 2.20), decreasing the bulk voltage attracts more holes from the source into the channel, effectively decreasing its resistance

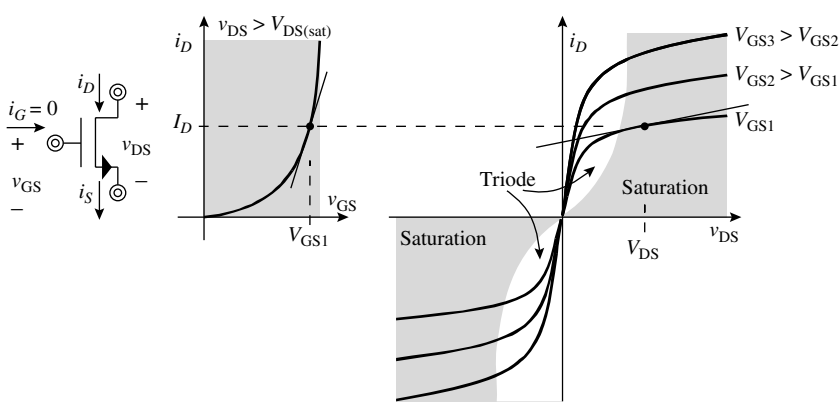


Figure 2.19 I-V curves for an n-channel MOSFET.

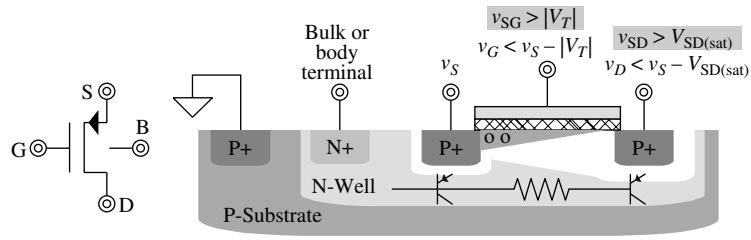


Figure 2.20 P-channel MOSFET in an n-well p-substrate CMOS technology.

and increasing current flow. Generally, forward biasing the source-bulk junctions of both p- and n-channel devices induces higher current flow, and vice versa. In essence, the bulk voltage changes the work function above which the gate voltage must surpass the source voltage to invert the region immediately beneath the gate and its effects are therefore accounted in threshold voltage v_T:

$$v_{TN} = V_{TNO} + \gamma(\sqrt{2\phi - v_{BS}} - \sqrt{2\phi}) \tag{2.25a}$$

and

$$|v_{TP}| = |V_{TPO}| + \gamma(\sqrt{2\phi - v_{SB}} - \sqrt{2\phi}) \tag{2.25b}$$

where v_{TN} and v_{TP} are the n- and p-channel threshold voltages, V_{TNO} and V_{TPO} the zero-bias n- and p-channel thresholds (when bulk-source voltage is zero), γ the bulk-effect parameter, and 2ϕ a process-dependent constant (normally 0.6 V).

The MOSFET is a capacitor-based transistor because it is through the oxide capacitance (and its electric field) that charge accrues directly beneath the gate. This is why parallel-plate oxide capacitors exist between gate and source and gate and drain, as can also be appreciated from the profile views of Figs. 2.17, 2.18, and 2.20. The gate must overlap the source and drain to ensure the inverted channel shorts them when in triode. As a result, a small overlap oxide capacitance C_{OV} exists between gate and source and gate and drain that is directly proportional to channel width W, overlap length L_{OV}, and oxide capacitance per unit area C''_{OX} (where $C_{OV} = WL_{OV}C''_{OX}$). The reverse-biased junctions attached to the drain and source introduce additional parasitic depletion-mode capacitors C_{SB} and C_{DB}:

$$C_{SB} = \frac{AC''_{jsbo}}{\sqrt{1 + \dfrac{v_{SB}}{V_{BI}}}} \tag{2.26}$$

$$C_{DB} = \frac{AC''_{jdbo}}{\sqrt{1 + \frac{v_{DB}}{V_{BI}}}} \tag{2.27}$$

where C''_{jsb0} and C''_{jdb0} are the zero-bias source- and drain-bulk capacitances per unit area. As before, the peripheral sidewall contributes additional and similar capacitance.

The main capacitance associated with the MOSFET is from gate to channel, but because the actual length of the channel varies with v_{DS} (Fig. 2.18), its value also changes with v_{DS}. In the triode region, for instance, when the length of the channel is the distance from source to drain (drawn length L), the total gate-channel capacitance splits equally between source and drain and is directly proportional to W, drawn channel length L, and C''_{OX} (i.e., $C_{GS} \approx C_{GD} \approx 0.5\ WLC''_{OX} + WL_{OV}C''_{OX}$). In the saturation region, however, the channel length shortens and the channel extends the source further, effectively increasing C_{GS} by another fraction of the channel capacitance (i.e., $C_{GS} \approx 0.67\ WLC''_{OX} + WL_{OV}C''_{OX}$) and decreasing C_{GD} to its overlap component (i.e., $C_{GD} \approx WL_{OV}C_{OX}''$). In the off region, both capacitors are relatively small and approximately equal to their overlap values (i.e., $C_{GS} \approx C_{GD} \approx WL_{OV}C''_{OX}$). Figure 2.21 illustrates graphically how these capacitances change across the operating modes of a MOSFET.

MOSFETs normally undergo a V_T-implant process step to ensure their threshold voltages are within acceptable window limits across process and temperature corners (e.g., 0.5–0.7 V). Devices not exposed to this implant adjustment are said to be *natural* or *native*. Natural NMOS transistors in p-type substrates have a nominal threshold voltage of approximately 0 V and natural PMOS devices −1.5 V.

To distinguish these natural devices and other transistors from one another, it is helpful to use distinctive schematic symbols, somehow illustrating their distinguishing electrical properties in a graphical, but subtle manner. The most basic three-terminal, capacitor-coupled gate MOS symbol belongs to the substrate, *enhancement-mode transistor* ($V_{TN} > 0$

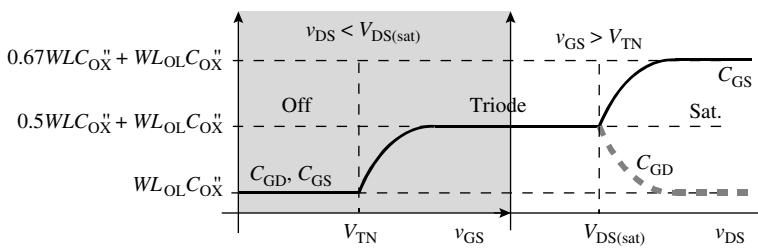


Figure 2.21 Gate-source and -drain capacitances across MOSFET operating regions.

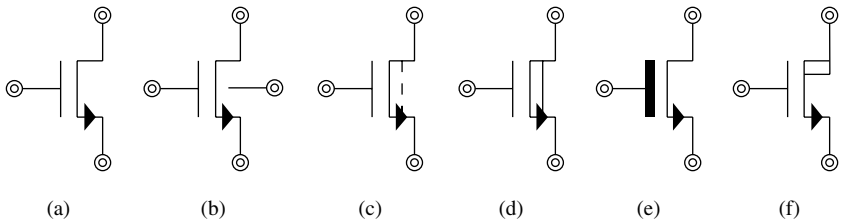


Figure 2.22 Typical symbols for (*a*) substrate, (*b*) bulk-isolated, (*c*) zero-V_T (e.g., natural NFET) (*d*) depletion-mode ($V_{TN} < 0$ V), (*e*) poly-2, and (*f*) drain-extended (i.e., n-well drain) NMOS transistors.

and $V_{TP} < 0$), as is the case for a substrate NMOS in a p-type substrate (Figs. 2.18 and 2.22*a*). A fourth back-gate terminal (Fig. 2.22*b*) indicates the bulk is isolated from its substrate, as exemplified by the n-well p-channel device shown in Fig. 2.20. An additional dashed line between source and drain (Fig. 2.22*c*) normally indicates the existence of a slightly inverted channel with zero gate-source voltages, as with natural NMOS devices whose threshold voltages are near zero. Similarly, an additional solid line between drain and source (Fig. 2.22*d*) typically represents the existence of a strongly inverted channel with zero gate-source voltages, as is the case for devices expressly V_T-adjusted for this purpose, otherwise known as *depletion-mode transistors*. A thicker gate line (Fig. 2.22*e*), on the other hand, may indicate the gate is built with a second-level polysilicon strip (with a thicker gate oxide between gate and channel), which is useful for applications requiring higher gate-source breakdown voltages. Although not as common, a modified drain line, as shown in Fig. 2.22*f*, may also indicate the drain has lower doping concentration, as with substrate NMOS transistors with n-well drains, which are useful in applications requiring higher drain-gate and drain-source breakdown voltages. These and other variations (combined or otherwise) also apply to PMOS transistors.

Extrinsic Device

The extrinsic portions of the MOSFET are the ohmic-series resistances and parasitic pn junctions attached to each terminal. A substrate n-channel MOSFET, for instance, as shown in Fig. 2.18, has parasitic source, drain, gate, and bulk resistances R_S, R_D, R_G, and R_B, respectively, and reverse-biased bulk-source and -drain pn-junction diodes (Fig. 2.23*a*). A welled MOSFET also has parasitic bipolar devices attached to its drain and source, as exemplified by the n-well p-channel MOSFET shown in Fig. 2.20 and modeled in Fig. 2.23*b*. These parasitic substrate bipolar devices channel current directly into the substrate, generating noise and therefore compromising the fidelity of the entire chip. This is the reason why designers normally avoid forward biasing the bulk-source junction, even though it may decrease the effective threshold voltage of the device.

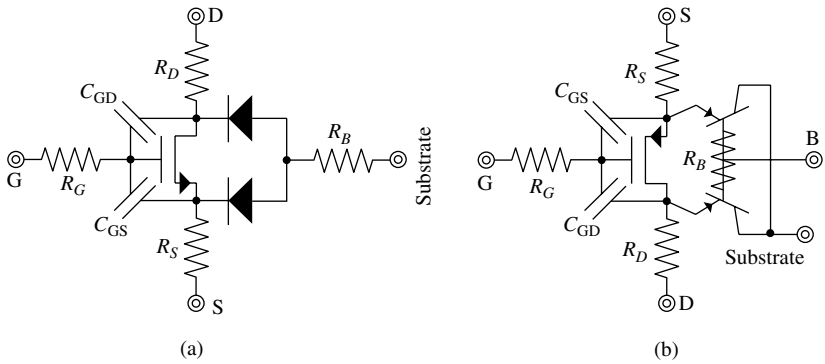


Figure 2.23 Large-signal models of (*a*) substrate n-channel and (*b*) n-well p-channel MOSFETs.

2.4.2 Small-Signal Response

Small-signal variations in all terminal voltages cause small-signal variations in the drain-source current that can be approximated reasonably well with linear, first-order models. More specifically, gate-, bulk-, and drain-source voltage variations v_{gs}, v_{bs}, and v_{ds}, respectively, induce small variations in drain current i_d via gate-channel capacitance, bulk, and channel-length modulation effects (Fig. 2.24*a*).

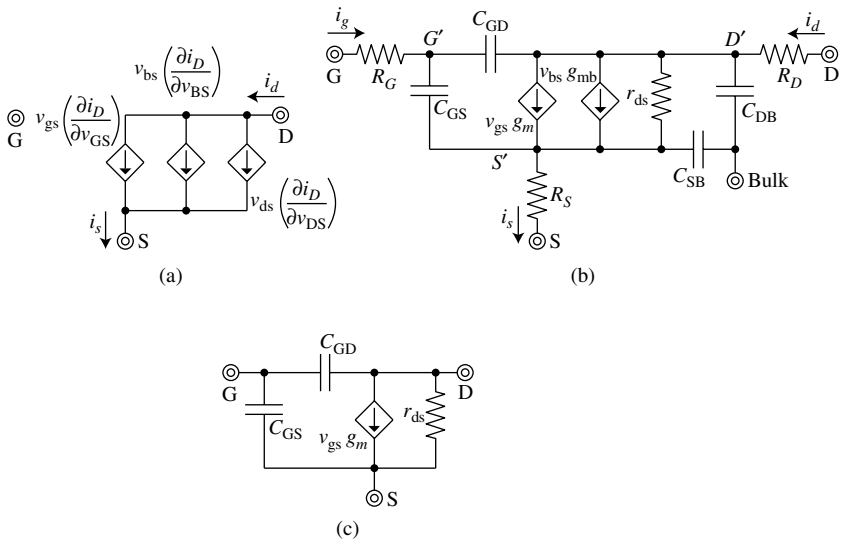


Figure 2.24 (*a*) Current-variation, (*b*) complete, and (*c*) simplified small-signal models of a MOSFET.

Consequently, the product of the slope of i_D with respect to v_{GS} at its biasing point (i.e., first partial derivative of i_D with respect to v_{GS}: $\partial i_D/\partial v_{GS}$) and small variation v_{gs} describes the resulting change in i_d:

$$i_d\Big|_{v_{ds}=0,v_{bs}=0} \approx v_{gs}\left(\frac{\partial i_D}{\partial v_{GS}}\right) \equiv v_{gs}g_m \approx v_{gs}\sqrt{2I_D K'\left(\frac{W}{L}\right)} \tag{2.28}$$

where g_m, as also shown in Fig. 2.24*b*, is the effective transconductance and I_D the biasing dc current of the MOSFET. Similarly, the product of the slope of i_D with respect to v_{DS} at its biasing point (i.e., first partial derivative of i_D with respect to v_{DS}: $\partial i_D/\partial v_{DS}$) and small variation v_{ds} describes the resulting change in i_d:

$$i_d\Big|_{v_{gs}=0,v_{bs}=0} \approx v_{ds}\left(\frac{\partial i_D}{\partial v_{DS}}\right) \equiv v_{ds}g_o \approx v_{ds}(\lambda I_D) \equiv \frac{v_{ds}}{r_{ds}} \tag{2.29}$$

where g_o is the effective output conductance of the device. Because this linear current variation results from a change in the voltage across its own two terminals v_{ds}, output resistor r_{ds}, as shown in Fig. 2.24*b*, with an equivalent resistance of $1/\lambda I_D$ can be used in place of dependent conductance g_o. Finally, the product of the slope of i_D with respect to v_{BS} at its biasing point (i.e., first partial derivative of i_D with respect to v_{BS}: $\partial i_D/\partial v_{BS}$) and small variation v_{bs} describes the resulting change in i_d, which is really the effect of modulating effective threshold voltage v_T and has the inverse effect of modulating v_{GS}:

$$\begin{aligned} i_d\Big|_{v_{ds}=0,v_{gs}=0} &\approx v_{bs}\left(\frac{\partial i_D}{\partial v_{BS}}\right) = v_{bs}\left(\frac{\partial i_D}{\partial v_T}\right)\left(\frac{dv_T}{dv_{BS}}\right) \\ &= v_{bs}\left(-\frac{\partial i_D}{\partial v_{GS}}\right)\left(\frac{dv_T}{dv_{BS}}\right) \\ &\approx -v_{bs}g_m\left(\frac{-\gamma}{2\sqrt{2\phi - V_{BS}}}\right) = v_{bs}\left(\frac{\gamma g_m}{2\sqrt{2\phi - V_{BS}}}\right) \equiv v_{bs}g_{mb} \end{aligned} \tag{2.30}$$

where V_{BS} is the dc bias bulk-source voltage and g_{mb} the effective transconductance, which is normally about one-tenth of the value of g_m. The effects of extrinsic devices R_G, R_S, R_D, C_{SB}, and C_{DB} are normally negligibly small, when considering first-order effects, and g_{mb} is often inconsequential because v_{BS} is frequently zero, which is why the simplified small-signal model of Fig. 2.24*c* is often useful.

Gate-source and gate-drain capacitances C_{GS} and C_{GD} shunt gate input signal v_{gs} at higher frequencies and consequently decrease the effective transconductance gain of the device. The frequency when short-circuit current gain i_d/i_{in} falls to one is defined as the transitional frequency f_T of the device beyond which the transistor is incapable of amplifying ac signals. During short-circuit conditions, when small-signal v_{ds} is zero, C_{GS} and C_{GD} are in parallel (neglecting parasitic ohmic resistors R_G, R_D, and R_S in Fig. 2.24*b*) and the voltage across them (i.e., v_{gs}) is the product of input current i_{in} and its equivalent parallel impedance,

$$\frac{i_d}{i_{in}} = \frac{v_{gs}g_m}{i_{in}}\bigg|_{v_d=0} = \left[\frac{i_{in}}{(C_{GS}+C_{GD})s}\right]\left(\frac{g_m}{i_{in}}\right)\bigg|_{f_T=\frac{g_m}{2\pi(C_{GS}+C_{GD})}} \equiv 1 \qquad (2.31)$$

where i_d/i_{in} is evaluated at its transitional frequency f_T. Higher transconductances with smaller device areas (i.e., smaller parasitic capacitance) yield higher transitional frequencies and are therefore deemed better (i.e., higher frequency) devices, which is why trench-isolated MOSFETs outperform their junction-isolated counterparts, because no parasitic capacitance from a common source-bulk terminal to substrate exists.

2.4.3 Layout

MOS devices match best when placed near one another, in the same orientation, with a common center of mass, cross-coupled, and with dummy devices around them to mitigate the effects of polyetching errors. The gate must extend beyond the width of the drain and source to ensure the channel, when inverted, fully extends across the entire width of the device, as shown in Fig. 2.25*a*. When bulk and

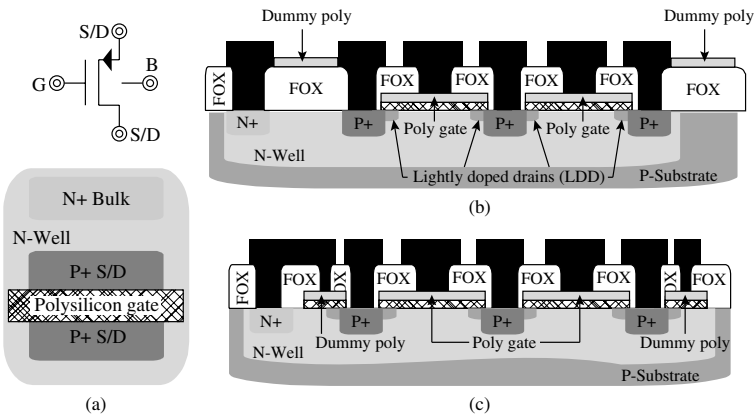


Figure 2.25 (*a*) Top-layout view of a p-channel MOSFET and relevant side profile views of matching lightly doped-drain (LDD) p-channel MOSFETs with (*b*) dummy and (*c*) active dummy polysilicon strips.

source are at the same potential, the n+ bulk region is normally butted against one of the p+ source/drain regions, forcing the latter to be the source.

To save space and circumvent polyetching errors, dummy polysilicon strips may be used in place of dummy transistors. These strips work reasonably well when the thin oxide regions underneath the polysilicon gates are on the order of 450 Å in height. Modern technologies, however, have oxide regions on the order of 125 Å or less, topographically placing the active polysilicon gates well beneath their peripheral dummy strips (Fig. 2.25*b*) and decreasing the extent to which the strips reduce etching errors. Plain polysilicon strips are automatically placed above thicker field oxide (FOX) regions to decrease their parasitic capacitance to the silicon substrate. It is therefore best to use dummy transistors, instead of plain polysilicon strips, biased in the off region or active dummy polysilicon gate strips, as shown in Fig. 2.25*c*. The active dummy strips may create parasitic MOS transistors and should therefore be laid out and biased to avoid inverting a channel immediately beneath them. For instance, an active dummy strip on an n-well region should be tied to the most positive potential to accumulate the region beneath them with majority n-type charge carriers.

With a thinner oxide, the electric field strength near the drain region increases (i.e., large drain-gate voltage across a relatively thin dielectric), compromising the reliability of the device and prematurely inducing impact-ionization and hot-carrier effects. To prevent these conditions, extending the drain with a *lightly doped drain* (LDD) region, as shown in Fig. 2.25*b* and *c*, reduces the charge carrier concentration and therefore decreases the probability of hot carriers crossing the gate oxide. Similarly, replacing the highly doped drain diffusion with a lighter doped well increases the drain-bulk breakdown voltage—these devices are sometimes known as *drain-extended* transistors.

2.5 Junction Field-Effect Transistors

2.5.1 Large-Signal Operation

Intrinsic Device

A *junction field-effect transistor* (JFET) is a diffusion resistor whose reverse-biased pn junctions "pinch" its conductive medium. Figure 2.26*a* and *b* illustrates how reverse-biased substrate-well and p+-well depletion regions pinch an n-well resistor (i.e., n-channel JFET), increasing its resistance as the channel pinches. The channel resistance is therefore directly proportional to channel length L and inversely proportional to channel width W, channel depth (i.e., as channel is less pinched with increasing gate-source voltage v_{GS}), channel doping

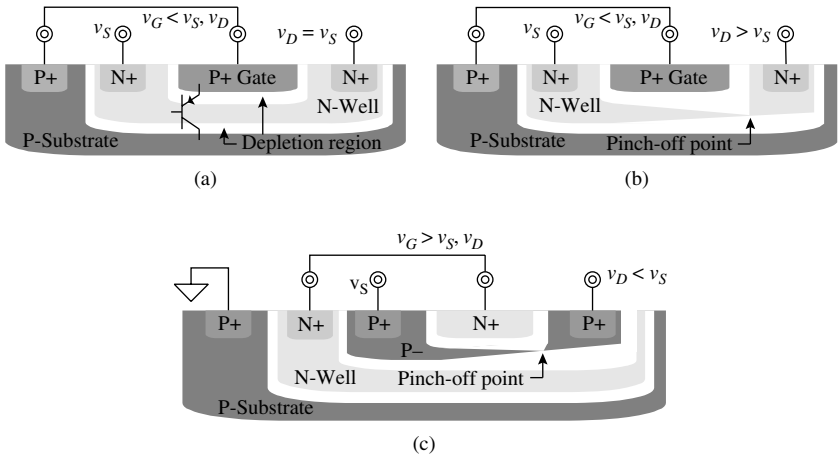


FIGURE 2.26 Side-profile views of a substrate n-channel junction field-effect transistor (NJFET) in (*a*) triode and (*b*) saturation (pinch off) and (*c*) a gate-isolated p-channel JFET in saturation (pinch off).

density, and majority charge-carrier mobility in the channel. The resulting drain current is the ratio of drain-source voltage v_{DS} and channel resistance $R_{Channel}$:

$$i_D = \frac{v_{DS}}{R_{Channel}} = v_{DS}\left(\frac{2I_{DSS}}{|V_P|^2}\right)[(v_{GS} + |V_P|) - 0.5v_{DS}]\Big|_{v_{DS}<V_{DS(sat)}}$$

$$\approx v_{DS}\left(\frac{2I_{DSS}}{|V_P|^2}\right)(v_{GS} + |V_P|) \tag{2.32}$$

where v_{DS} is low (i.e., in triode, when v_{DS} is below saturation voltage $V_{DS(sat)}$) and the ratio of I_{DSS} and pinch-off voltage squared V_P^2 is directly proportional to W, doping density, and mobility and inversely proportional to L. When v_{DS} increases beyond the channel pinch-off point (i.e., v_{DG} is greater than $|V_P|$ or v_{DS} greater than $V_{DS(sat)}$):

$$V_{DS(sat)} = V_{GS} + |V_P| \tag{2.33}$$

The voltage across the channel and (therefore) drain current i_D remain relatively constant, irrespective of v_{DS}, except the effective length of the channel and (consequently) channel resistance $R_{Channel}$ decrease slightly with increasing v_{DS},

$$i_D\Big|_{v_{DS}>V_{DS(sat)}} = \frac{V_{DS(sat)}}{R_{Channel}} \approx \left(\frac{I_{DSS}}{V_P^2}\right)(v_{GS} + |V_P|)^2(1 + \lambda v_{DS}) \tag{2.34}$$

where λ is the channel-length modulation parameter and the device is said to be saturated or in *pinch off*.

The JFET is a drift-based device and its drain-current relationships therefore mimic those of the MOSFET, including its I-V curves (Fig. 2.19). The driving difference between the two, from an operational standpoint, is that the JFET is normally on with zero gate-source voltages and the MOSFET is typically not. The JFET also has a small gate current (i.e., the reverse-saturation current of its gate-channel pn junctions), which is nonexistent in the MOS counterpart. Because the channel dimensions depend on the out-diffusion lengths of the channel and gate (instead of the photolithographic resolution of processing masks), shrinking the JFET to MOSFET levels, following Moore's law, is not as straightforward. As a result, considering the demand for increasingly dense system-on-chip (SoC) integration, MOSFETs enjoy the popularity JFETs cannot hope to achieve. However, because the channel is well beneath the gate, away from surface irregularities and imperfections, the device exhibits less $1/f$ noise, finding a niche market in low-noise applications. Additionally, since JFET resistances can be substantially large, designers often use them as high-value resistors (i.e., as *pinched resistors*).

The JFET is essentially a resistor with two reverse-biased pn junctions. As such, its schematic symbol is a solid line between drain and source, denoting the explicit channel resistor, with an arrow pointed at or away from the top gate that is capacitively coupled to the channel via a pn-junction depletion-mode capacitor to indicate the orientation of its pn-junction gate. The n-channel device has the arrow pointing to the gate (Fig. 2.27*a*) and two depletion-mode capacitors connected to source and drain,

$$C_{\text{Dep}} = \frac{AC''_{\text{jxo}}}{\sqrt{1 + \dfrac{v_R}{V_{\text{BI}}}}} \tag{2.35}$$

where gate-source- and -drain capacitances C_{GS} and C_{GD} in Fig. 2.27 conform to C_{Dep} above when C''_{jx0} is replaced with their respective zero-bias capacitances per unit area and v_R with their respective reverse-bias voltages (note peripheral sidewall contributes additional and similar capacitance). The channel resistance, as with most diffusion resistors, normally exhibits a positive temperature coefficient, becoming weaker (i.e., conducting less current) at higher temperatures.

Extrinsic Device

Series gate, source, and drain resistances R_G, R_S, and R_D, respectively, comprise the extrinsic portion of the device. There is also a parasitic vertical BJT in the substrate n-channel JFET with the source as its base, as shown in Fig. 2.26*a*, except all relevant pn junctions are reverse biased and the BJT is consequently in the off mode, that is, practically

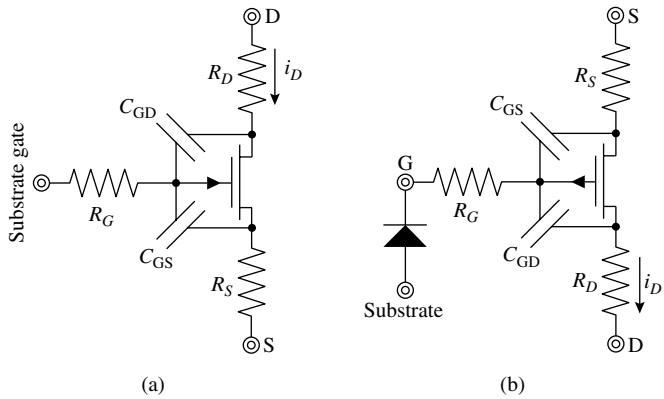


FIGURE 2.27 Complete large-signal models of (*a*) substrate n- and (*b*) gate-isolated p-channel JFETs.

nonexistent. Since the bottom gate of the p-channel JFET in Fig. 2.26*c* is a well immersed in the substrate, an extrinsic pn-junction diode to the substrate also exists, as modeled in Fig. 2.27*b*.

2.5.2 Small-Signal Response

Gate- and drain-source voltage variations v_{gs} and v_{ds}, respectively, induce small variations in drain current i_d via the pinching and channel-length modulation effects, both of which are reasonably modeled with linear, first-order approximations (Fig. 2.28*a*). The product of the slope of i_D with respect to v_{GS} at its biasing point (i.e., first partial derivative of i_D with respect to v_{GS}: $\partial i_D/\partial v_{GS}$) and small variation v_{gs} describes the resulting change in i_d:

$$i_d\big|_{v_{ds}=0} \approx v_{gs}\left(\frac{\partial i_D}{\partial v_{GS}}\right) \equiv v_{gs}g_m \approx v_{gs}\sqrt{4I_D\left(\frac{I_{DSS}}{V_P^2}\right)} \tag{2.36}$$

where I_D is the biasing dc current of the JFET. Similarly, the product of the slope of i_D with respect to v_{DS} (i.e., first partial derivative of i_d with respect to v_{ds}: $\partial i_D/\partial v_{DS}$) and small variation v_{ds} describes the resulting change in i_d:

$$i_d\big|_{v_{gs}=0} \approx v_{ds}\left(\frac{\partial i_D}{\partial v_{DS}}\right) \equiv v_{ds}g_o \approx v_{ds}(\lambda I_D) \equiv \frac{v_{ds}}{r_{ds}} \tag{2.37}$$

Because this last linear current variation results from a change in the voltage across its own two terminals v_{ds}, output resistor r_{ds} (Fig. 2.28*b*) with an equivalent resistance of $1/\lambda I_D$ can be used in place of g_o. The effects of extrinsic devices R_G, R_S, and R_D are normally negligibly small,

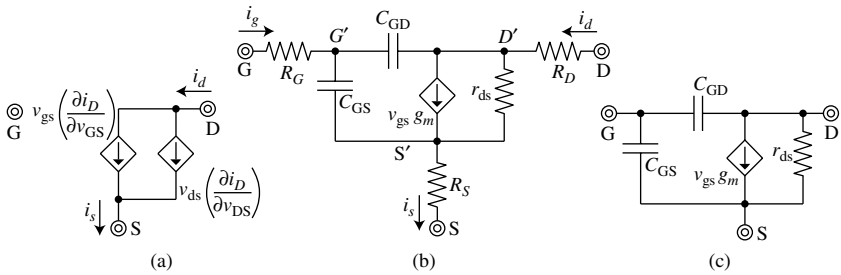


Figure 2.28 (*a*) Current-variation, (*b*) complete, and (*c*) simplified small-signal models of a JFET.

when considering first-order effects, which is why the simplified small-signal model of Fig. 2.28*c* is often useful.

Gate-source and gate-drain capacitances C_{GS} and C_{GD} shunt gate input signal v_{gs} at higher frequencies and consequently decrease the effective transconductance gain of the device. The frequency when short-circuit current gain i_d/i_{in} falls to one is defined as the transitional frequency f_T of the device beyond which the transistor is incapable of amplifying ac signals. During short-circuit conditions, when small-signal v_{ds} is zero, C_{GS} and C_{GD} are in parallel (when neglecting parasitic ohmic resistors R_G, R_D, and R_S) and the voltage across them (v_{gs}) is the product of input current i_{in} and its equivalent parallel impedance,

$$\frac{i_d}{i_{in}} = \frac{v_{gs}g_m}{i_{in}}\bigg|_{v_d=0} = \left[\frac{i_{in}}{(C_{GS}+C_{GD})s}\right]\left(\frac{g_m}{i_{in}}\right)\bigg|_{f_T=\frac{g_m}{2\pi(C_{GS}+C_{GD})}} \equiv 1 \qquad (2.38)$$

where i_d/i_{in} is evaluated at its transitional frequency f_T.

2.5.3 Layout

As with all previous devices, JFETs match best when placed near one another, in the same orientation, with a common center of mass, cross-coupled, and with dummy peripheral devices around them to mitigate out-diffusion mismatch effects. In the case of the substrate n-well JFET shown in Fig. 2.26*a* and *b*, the top p+ diffusion must extend beyond the width of the n-well to fully pinch the resistor (and avoid short-circuiting the resistor on the sides), as shown in Fig. 2.29*a*, in the process short-circuiting the gate to the substrate. Because the connectivity of the gate is limited to that of the substrate, the transistor is said to be a substrate device. In the case of the p-channel JFET illustrated in Fig. 2.26*c*, the bottom gate is isolated from the substrate so the gate terminal has more flexibility, but as with the substrate

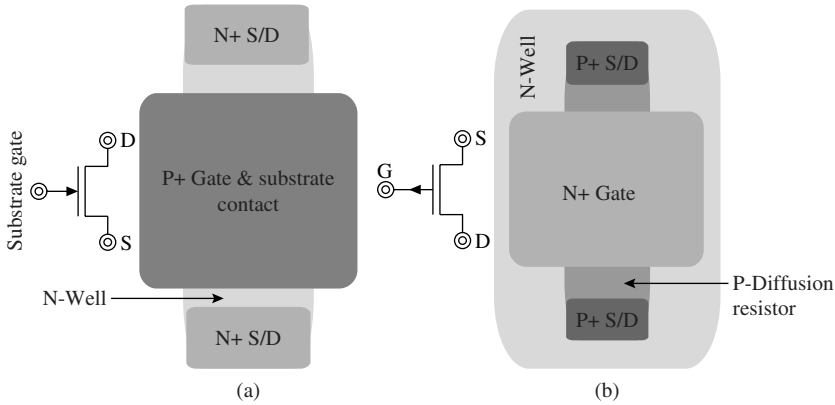


FIGURE 2.29 Top-layout views of (*a*) substrate n- and (*b*) gate-isolated p-channel JFETs.

JFET, its top gate must extend beyond the width of the resistor and overlap with the n-well bottom gate.

2.6 Summary

The resistance of resistors, as are MOSFETs and JFETs, depends on the physical characteristics and dimensions of the current-conducting medium, that is, on its resistivity, length, width, and depth. The capacitance of capacitors, on the other hand, depends on the surface area of the facing plates, the distance between them, and the permittivity of the material separating them. A diode, unlike resistors and capacitors, is a diffusion device that conducts current only when forward biased, and its relationship to its voltage is exponential. BJTs exploit this exponential diffusing-current relationship by exposing diffused charge carriers to a reverse-biased pn junction that is immediately next to the forward-biased device, sweeping all carriers to a collector terminal—note modulating the voltage or current into the middle region (i.e., the base) varies the current. MOSFETs and JFETs vary the channel resistance across its terminals via an electric field, inducing a change in current via drift effects. FETs are therefore voltage-driven devices, unlike the BJTs, which are current driven. Ultimately, with respect to application, small-signal changes in voltages induce approximately small *linear* variations in current. Small-signal variations, linear transconductances, input and output resistances, and capacitances are arguably the most important analog parameters of a device, as they describe how a device converts voltages into currents, and vice versa, across frequency, up to its transitional frequency point f_T beyond which signals can no longer be amplified.

Understanding how these devices work intrinsically is the first step in building robust ICs. Extrinsic parasitic effects, however, exist, and if not minimized through careful physical and electrical design, their impact on a system may be fatal. Ultimately, the intrinsic and extrinsic limits of the physical components comprising a system set the operational limits of the same, especially when exposed to process and temperature extremes. Chapter 3, as a follow-up step in the analog IC design process, discusses how to combine these devices to build the basic analog circuit blocks necessary to design analog subsystems like the voltage regulator.

CHAPTER 3
Analog Building Blocks

Integrated circuits (ICs) achieve prescribed functions by using microelectronic devices to process currents and/or voltages. The electronic systems they comprise vary in nature and often call for sensing and/or conditioning naturally existing signals that (1) have low- or high-energy content and (2) require voltage-voltage, current-voltage (transresistance), voltage-current (transconductance), and/or current-current conversion and amplification or attenuation (i.e., conditioning). Common analog applications include regulation and control, as in the case of linear regulators, where enclosing signal-processing amplifiers in closed feedback loops ensures the sensed outputs in the loop remain unchanged when confronted with a changing environment.

A regulator, in particular, senses, amplifies, and controls its conductance to ensure the output is within specified limits, irrespective of its load, input, and operating conditions. Sensing the output voltage, converting it to current, processing current, and converting it back to voltage are intrinsic in this process, which is why amplifiers, converters (of all sorts), and current mirrors constitute basic building blocks in analog ICs. Resistors, capacitors, and transistors, be they bipolar-junction transistors (BJTs), metal-oxide semiconductor field-effect transistors (MOSFETs), or junction field-effect transistors (JFETs), to this end, channel and convert currents into voltages, and vice versa, in the process amplifying and conditioning signals as desired and prescribed by the designer and the process technology used. This chapter discusses how microelectronic devices (alone and combined) achieve the basic analog functions necessary to design and build higher order analog systems. Subsequent chapters then discuss how these blocks combine to build and establish stable and robust regulating feedback loops, the ultimate purpose of which, as it pertains to this textbook, is to design high-performance low-dropout regulators.

3.1 Single-Transistor Amplifiers

3.1.1 Common-Emitter/Source Amplifiers

Single-transistor amplifiers are among the most basic, yet most useful circuits in analog IC design, especially the common-emitter (CE) and common-source (CS) transistors. CE and CS stages, as illustrated in Fig. 3.1*a*, have their emitters and sources connected to low-impedance points (i.e., ac grounds), which is why those terminals are said to be at *common* nodes. Because collector and drain currents are strong functions of base and gate voltages and weak functions of collector and drain voltages, bases and gates in CE and CS circuits are used as input terminals and collectors and drains as outputs. The collector or drain load is typically a combination of complementary p-type CE and CS stages, current sources, resistors, and/or whatever other impedances the surrounding electronics introduce.

Large-Signal Operation

CE and CS stages are voltage inverters because their outputs are high when their respective inputs are low, and vice versa. For instance, when the base-emitter or gate-source voltage (i.e., input voltage v_{IN}) in an n-type CE/CS stage is below the level necessary to sink load current I_{Bias} (i.e., v_{IN} is less than $V_{IN(high)}$ in the transfer function shown in Fig. 3.1*b*), the load pulls output v_{OUT} closer to the positive supply (i.e., toward $V_{OUT(max)}$). Similarly, when v_{IN} is above the level necessary to sustain I_{Bias} (i.e., v_{IN} is greater than $V_{IN(low)}$), the transistor pulls v_{OUT} closer to the negative supply (i.e., toward $V_{OUT(min)}$).

Because the n-type inverting device is off when v_{IN} is zero, the output reaches the positive supply (i.e., $V_{OUT(max)}$ is positive supply

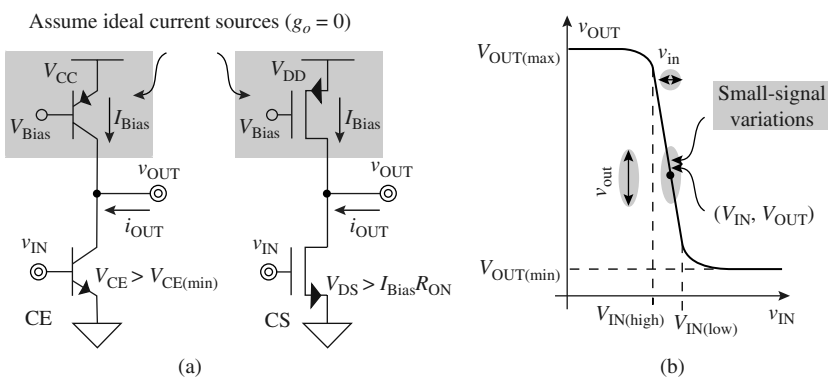


Figure 3.1 (*a*) Common-emitter (CE) and common-source (CS) gain circuits and (*b*) their large-signal Response.

voltage V_{CC} or V_{DD}). When the input is at its maximum driving voltage, however, the inverting transistor pulls the output closer to the negative supply, but not completely. Under maximum driving conditions, in fact, a BJT pulls the output within $V_{CE(min)}$ or 0.2–0.3 V of the lower supply because the base-collector pn-junction diode forward biases and clamps v_{OUT}. In the MOSFET case, the triode channel resistance determines the extent to which the transistor pulls v_{OUT} to its respective supply: $V_{OUT(min)}$ is within a $I_{Bias}R_{Channel.Triode}$ (or $I_{Bias}R_{ON}$) voltage from the supply. Note that allowing the BJT to enter the saturation region slightly has little to no impact on its high-gain characteristics because the base-collector junction current remains substantially lower than the base-emitter current, effectively extending its high-gain collector-emitter voltage range to within approximately 0.2–0.3 V of the supply. Conventionally, deep saturation limit $V_{CE(min)}$, beyond which the BJT undergoes significant changes, is said to be roughly 0.2–0.3 V.

Small-Signal Response at Low Frequency

In most analog applications, designers bias the amplifying transistor at a point in its large-signal range that produces the highest voltage gain. As such, input signal v_{IN} comprises dc biasing voltage V_{IN} and ac input voltage v_{in}, the former of which is used to ensure dc output voltage V_{OUT} is between the supplies (Fig. 3.1*b*) and the amplifying transistor is in its high-gain mode (i.e., forward active or slightly saturated for a BJT and saturation for a MOSFET). Under these quiescent biasing conditions, small ac variations in v_{IN} (i.e., v_{in}) generate relatively larger ac variations in v_{OUT} (i.e., v_{out}), as denoted by the relatively steep slope at biasing point (V_{IN}, V_{OUT}) in Fig. 3.1*b*.

To determine the low-frequency voltage gain of the circuit A_{V0}, the amplifying transistor is replaced with its small-signal equivalent circuit model, as shown in Fig. 3.2. For simplification, loading resistor and capacitor R_{Load} and C_{Load} collectively embody all loading effects, including those of bias current generator I_{Bias}. Low-frequency output voltage v_{out} is therefore the ohmic voltage drop across the parallel combination of output resistor r_o or r_{ds} and loading resistor R_{Load},

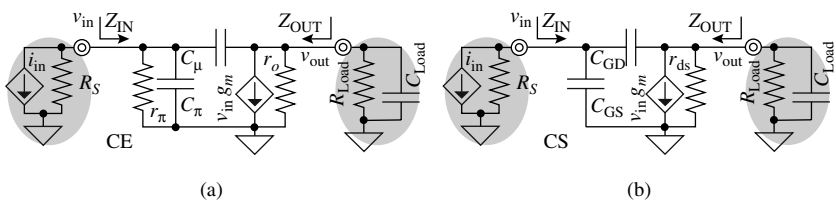


Figure 3.2 Small-signal models of the (*a*) common-emitter (CE) and (*b*) common-source (CS) amplifiers with their respective source- and load-equivalent circuits.

which results from pulling transconductor current $v_{be}g_m$ or $v_{gs}g_m$ (i.e., $v_{in}g_m$) through the resistors:

$$A_{V.\mathrm{LF}} \equiv \frac{v_{\mathrm{out}}}{v_{\mathrm{in}}} = \frac{-v_{\mathrm{in}}g_m(r_o \parallel R_{\mathrm{Load}})}{v_{\mathrm{in}}} = -g_m(r_o \parallel R_{\mathrm{Load}}) \tag{3.1}$$

where transconductance g_m can be micro- to milliamps per volt and output resistor r_o or r_{ds} kilohms to megohms. Unloaded gain $-g_m r_o$ or $-g_m r_{ds}$ (where R_{Load} is infinitely large) is said to be the *intrinsic* or *maximum gain* of the transistor, as connecting anything else to the output reduces its effective output resistance and therefore decreases its voltage gain. This relatively high voltage gain is, for the most part, the driving feature of the CE/CS transistor. Note CE and CS circuits invert their input signals by 180°.

Normally, a single stage does not provide sufficient gain and, in cascading several, two-port model equivalents of each successive stage are useful. The ac-equivalent circuits shown in Fig. 3.2 are simple enough that they also constitute useful two-port models for CE and CS stages. The three most important parameters associated with these models are small-signal input resistance R_{IN}, output resistance R_{OUT}, and transconductance gain G_M, which in this case are r_π or infinitely high, r_o or r_{ds}, and g_m, respectively. Input resistances can be 100–500 kΩ (or infinitely high in the MOS case), output resistances 0.5–5 MΩ, and transconductances 0.01–1 mS, yielding voltage gains between 0.1–1 kV/V or 40–60 dB.

At times, CE and CS stages implement transconductance functions with their collector or drain currents as their main output-processed signals. In these cases, since the output signal is a current, the output resistance of the circuit, like in a current source, should be substantially high. Output resistances in the 0.5–5 MΩ range, however, may not be high enough, which is why a resistor is sometimes inserted between the emitter/source and the low-impedance supply, as shown in Fig. 3.3, to increase the overall series resistance of the circuit. To ascertain this new output resistance, the input is connected to ac ground (i.e., v_{in} is zero) and a test voltage applied to the output. Because gate and bulk are both at ac grounds and g_m and g_{mb} are consequently in parallel and both directly proportional to source voltage v_s (since $v_{gs} = -v_s$), the effective transconductance of the MOS case for this analysis is the sum of g_m and g_{mb}. The ratio of the output voltage to its resulting current represents the short-circuit small-signal output resistance of the circuit:

$$\begin{aligned} R_{\mathrm{OUT}}\Big|_{v_{\mathrm{in}}=0} &\equiv \frac{v_{\mathrm{out}}}{i_{\mathrm{out}}} = \frac{(i_{\mathrm{out}} - v_{gs}g_{m.\mathrm{eff}})r_{ds} + v_s}{i_{\mathrm{out}}} \\ &= \frac{(i_{\mathrm{out}} + Ri_{\mathrm{out}}g_{m.\mathrm{eff}})r_{ds} + Ri_{\mathrm{out}}}{i_{\mathrm{out}}} \\ &= r_{ds} + R(g_m + g_{\mathrm{mb}})r_{ds} + R \approx R(g_m + g_{\mathrm{mb}})r_{ds} \end{aligned} \tag{3.2}$$

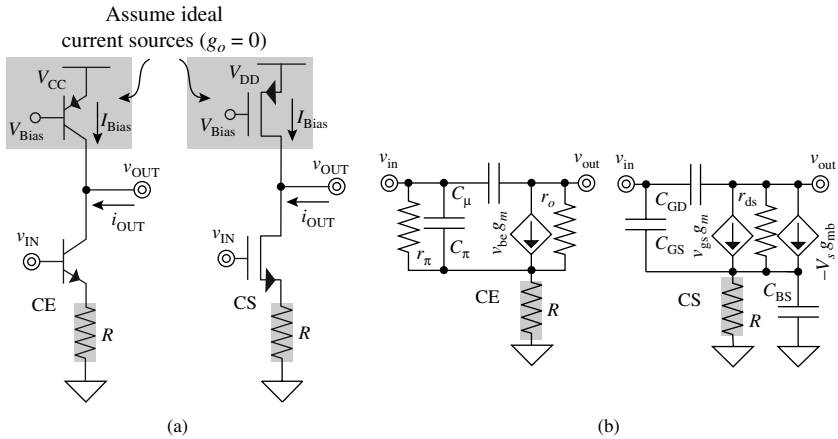


FIGURE 3.3 (*a*) Emitter- and source-degenerated common-emitter and -source gain stages and (*b*) their respective small-signal equivalent circuits.

where v_{out} is the test voltage applied across the output. Because $g_m r_{ds}$ can be 40–60 dB, the $R(g_m + g_{mb})r_{ds}$ term normally overwhelms r_{ds} and R, reducing the above expression to $R(g_m + g_{mb})r_{ds}$, which again, is considerably larger than the original output resistance of a CS circuit (i.e., r_{ds}). Bulk-effect parameter g_{mb}, by the way, disappears if the bulk is connected to the source. In the case of a BJT, bulk effects are also nonexistent (i.e., g_{mb} disappears) and, because v_{in} is zero, r_π is in parallel with R, shunting R (i.e., $R \,||\, r_\pi$) and reducing its output resistance to $(R \,||\, r_\pi)g_m r_o$ or βr_o, if R is considerably larger than r_π:

$$R_{OUT}\Big|_{v_{in}=0} \equiv \frac{v_{out}}{i_{out}} = r_o + (R||r_\pi)g_m r_o + (R||r_\pi) \approx (R||r_\pi)g_m r_o \le \beta r_o \quad (3.3)$$

Unfortunately, inserting R into the input-signal path also affects the transconductance gain of the amplifier. Only a fraction of the input signal is now impressed across the base-emitter or gate-source terminals, decreasing the short-circuit transconductance current flowing through the transistor by a factor of $(1 + Rg_m)$:

$$G_{M.LF}\Big|_{v_{out}=0} = \frac{i_{out}}{v_{in}} \approx \frac{v_{gs}g_m}{v_{gs} + v_R} = \frac{v_{gs}g_m}{v_{gs} + R(v_{gs}g_m)} = \frac{g_m}{1 + Rg_m} \le \frac{1}{g_m} \quad (3.4)$$

Assuming the currents flowing through r_{ds} and g_{mb} are negligibly smaller than $v_{gs}g_m$, where $G_{M.LF}$ is the effective low-frequency transconductance of the circuit evaluated when v_{out} is zero to ensure $G_{M.LF}$ does not model what is already modeled by R_{OUT}, v_R (or v_s) is the

voltage across R and $v_{gs}g_m$ is the current flowing through R. The resulting decrease in transconductance, now that $G_{M.LF}$ is less than g_m, is the reason why series resistor R is said to be an *emitter-* or *source-degenerating resistor* and $G_{M.LF}$ the *degenerated transconductance* of the CE/CS amplifier. In fact, when the degeneration voltage across R is sufficiently large, $G_{M.LF}$ reduces to $1/R$. Degenerated CE/CS transistors, by the way, are also known as *cascode* devices, especially when their degenerating resistors are transistors.

Bulk effects also degenerate $G_{M.LF}$ because g_{mb} is in parallel with g_m and decreases with increasing v_R. Unlike g_m, however, the current does not increase with the input gate voltage so it has no impact on the "generating" numerator of $G_{M.LF}$, just the "degenerating" denominator:

$$G_{M.LF}\Big|_{v_{out}=0} = \frac{i_{out}}{v_{in}} \approx \frac{v_{gs}g_m - v_R g_{mb}}{v_{gs} + v_R} = \frac{g_m}{1+(g_m + g_{mb})R} \le \frac{1}{g_m} \tag{3.5}$$

where v_R or v_s is $i_{out}R$ and i_{out} is the difference between $v_{gs}g_m$ and $v_R g_{mb}$. In the BJT case, bulk effects disappear, and although the result is similar, an additional base current i_b flows through R that is a $1/\beta$ translation of collector transconductor current $v_{be}g_m$:

$$G_{M.LF}\Big|_{v_{out}=0} = \frac{i_{out}}{v_{in}} \approx \frac{v_\pi g_m}{v_\pi + v_R} = \frac{v_\pi g_m}{v_\pi + R\left[v_\pi g_m\left(1+\frac{1}{\beta}\right)\right]}$$

$$\approx \frac{g_m}{1+Rg_m} \le \frac{1}{g_m} \tag{3.6}$$

The additional series resistance also increases the input resistance (R_{IN}) of the circuit, except in the MOS case where input resistance is already infinitely high:

$$R_{IN} = \frac{v_{in}}{i_{in}} = \frac{v_\pi + v_R}{i_b} \approx \frac{i_b r_\pi + (1+\beta)i_b R}{i_b} = r_\pi + (1+\beta)R \tag{3.7}$$

where the total current flowing through R is base current i_b and collector current i_c (i.e., βi_b). Ultimately, because R_{OUT} is larger than r_o or r_{ds}, the total output resistance of the circuit, including its load, normally reduces to R_{Load} and, since $G_{M.LF}$ is lower than g_m, low-frequency voltage gain $A_{V.LF}$ tends to be lower than its nondegenerated counterpart:

$$A_{V.LF} \equiv \frac{v_{out}}{v_{in}} = \frac{-v_{in}G_{M.LF}(R_{OUT} \parallel R_{Load})}{v_{in}} \approx \frac{-g_m(R_{Load})}{1+Rg_m} \tag{3.8}$$

Physical Interpretation of Poles and Zeros

To start, the parallel combination of a resistor and a capacitor constitutes a pole:

$$\left(R \parallel \frac{1}{sC}\right) = \frac{R}{1+sRC} = \frac{R}{1+\frac{2\pi s}{p}}$$

From control theory, as shown in Fig. 3.4*a*, poles decrease the gain of a circuit past their respective locations at a rate of 20 dB per decade of frequency and lose a total of 90° of phase per pole. This reduction in gain, in physical terms, corresponds to a signal losing energy through a shunting capacitor and manifests itself as a lower root-mean-square (RMS) current and voltage. Consider, for example, how input and output capacitors C_{Pin} and C_{Po} in Fig. 3.4*b* steer energy (RMS current) away from input v_{in} and output v_o. The current and energy left for input and output resistances R_{IN} and R_O are lower (and so are their respective voltages), producing the pole effects shown in the Bode plot of Fig. 3.4*a*. Because the impedance through the capacitor ($1/sC$) decreases linearly with frequency, so does the corresponding gain of a circuit in the presence of a pole, which is why there is a factor of ten change in gain (i.e., 20 dB) for every tenfold increase in frequency (i.e., every decade of frequency). Incidentally, feed-forward capacitors $C_{\text{FF.LHP}}$ and $C_{\text{FF.RHP}}$ in Fig. 3.4*b* and *c* also steer energy away from v_{in} and similarly add to the shunting effects of C_{Pin} to produce the pole effects just described.

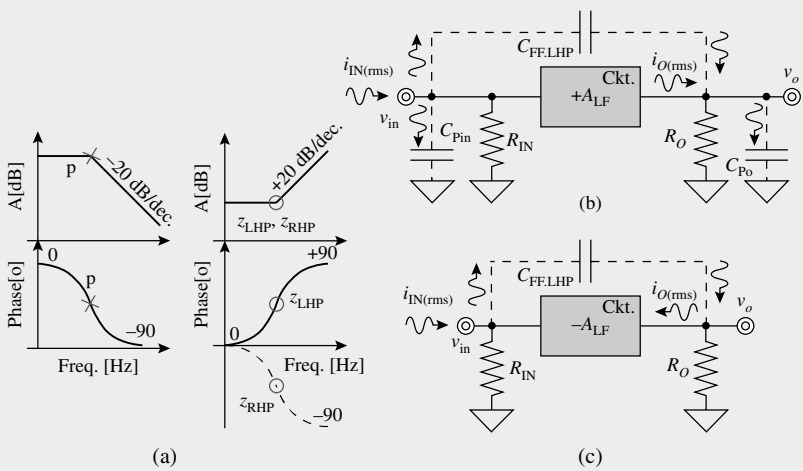


Figure 3.4 (*a*) Gain and phase Bode-plot responses of left-half-plane poles (e.g., p_{IN} and p_o) and zeros (e.g., z_{LHP}) and right-half-plane zeros (e.g., z_{RHP}) and (*b*) and (*c*) the circuits that produce them.

What is interesting about feed-forward capacitors $C_{FF.LHP}$ and $C_{FF.RHP}$ is that they add energy to the output, which is the opposite effect of poles. Recall from control theory, as shown in the figure, a zero increases the gain past its location, which is exactly the effect feed-forward capacitors have on a signal, given they add more energy to a signal as their impedances decrease with frequency. Since designers often rely in the number of inversions across a circuit to stabilize negative feedback loops, placing feed-forward capacitors (e.g., $C_{FF.LHP}$) across noninverting gain stages, as shown in Fig. 3.4*b*, keeps the integrity of signal inversions across the circuit. This type of zero is useful and corresponds to a left-half-plane zero whose net effect is to add 90° of total phase. A feed-forward capacitor that carries the opposite polarity of the signal present at low frequencies like $C_{FF.RHP}$, as shown in Fig. 3.4*c*, undermines the integrity of the low-frequency negative feedback loop because it has a tendency to invert the phase of the output. Although this zero still adds energy to the output and the gain increases, the phase shift across the circuit *decreases* a total of 90°, like a pole, unlike a conventional left-half-plane zero.

Another way to interpret pole-zero effects in a circuit is to consider impedance. Consider, for example, the RC network shown in Fig. 3.5*a*. At low frequencies, the capacitor is an open circuit and the resistance is at its maximum value, so the transresistance gain at that point is also high, as shown in the frequency response of Fig. 3.5*b*. As frequencies increase, past the point the impedance across capacitor C is equal to R_2, the total impedance from signal v_s to ground decreases and so does its gain—this drop in gain as frequency increases corresponds to a pole. At higher frequencies, however, when C's impedance is substantially lower than R_1, the total resistance across the network reduces to R_1, which, with respect to frequency, flattens the gain. This flattening effect, because it opposes the effects of the preceding pole, corresponds to a zero—this zero is on the left-half plane because there was no signal inversion across the circuit.

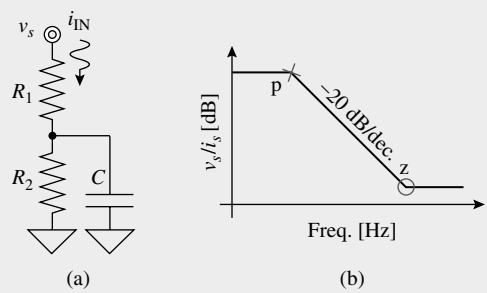


FIGURE 3.5 Changing impedance effects of (*a*) an RC network and (*b*) its corresponding frequency response.

Time-Constant Technique

In ascertaining the response of bandwidth-limited, or more to the point, low-pass filter circuits across frequency, it is often useful to determine the location of dominant, low-frequency poles using the time-constant (TC) technique. The basis of the foregoing approach is that the first capacitor to short-circuit with respect to its parallel resistance sets the dominant, low-frequency pole of a circuit. The TC analysis approximates this dominant pole by looking at every capacitor in the circuit independently, while all other capacitors are open-circuited, and the one that produces the highest RC time constant (i.e., first capacitor to short) corresponds to the capacitor introducing the lowest frequency pole (i.e., first pole as frequency increases: p_1):

$$p_1 \approx \frac{1}{2\pi \sum_{i=1}^{N} (R_i C_i)}\Bigg|_{R_x C_x >> R_i C_i} \approx \frac{1}{2\pi R_x C_x} \tag{3.9}$$

where N is the total number of capacitors, i is evaluated from 1 to N, time constant R_xC_x is considerably larger than all other time constants, and C_x is the dominant pole-setting capacitor. Note parallel resistance R_i is with respect to the capacitor and not ac ground, which is why, when performing this type of analysis, it is helpful to think of one of the capacitor terminals as "virtual ground" and ac grounds simply as other nodes in the circuit.

The TC method is even more powerful when applied successively to determine the location of higher frequency poles. For instance, applying the TC test to a circuit whose dominant, low-frequency pole-setting capacitor is short-circuited yields the location of the second pole, the next low-frequency pole as frequency increases. Again, short-circuiting the capacitor associated with the second pole while keeping the first one short-circuited and applying the TC approach yields the location of the third pole, the next pole in the sequence. In general, modifying (and simplifying) the circuit to accommodate for each successive capacitor that short-circuits produces new frequency-dependent circuits against which the TC technique can be applied. This method of approximating poles is more accurate when the frequency spread between poles is relatively wide (i.e., $p_1 << p_2 << \ldots << p_N$) because the capacitor impedances of closely spaced poles affect one another's time constants and their independent analysis becomes less accurate as a result.

Small-Signal Response at High Frequency

Base-collector C_μ and gate-drain C_{GD} capacitors have peculiar effects on gain. To start, relative to their inverting gain paths, C_μ and C_{GD} constitute parallel, out-of-phase, feed-forward paths because, as the capacitors short-circuit at higher frequencies, they carry more of the input signal onto its output, but opposite in phase to the transconductor current. A parallel feed-forward path indicates the presence of a zero because v_{out} receives more of the input signal and therefore increases in magnitude. An out-of-phase path implies the zero is on the right-half plane because the polarity opposes that of the main path, as dictated by the transconductor current. These effects only occur at higher frequencies, however, when the reversing ac current through C_μ or C_{GD} exceeds transconductor current i_{gm} (i.e., $v_{be}g_m$ or $v_{gs}g_m$), and during the transitional point, when the currents equal, no current flows through r_o or r_{ds} and v_{out} is consequently zero:

$$i_{C\mu} = (v_{be} - v_{ce})sC_\mu = (v_{be} - 0)sC_\mu\Big|_{z_{RHP}=\frac{g_m}{2\pi C_\mu}} \equiv i_{gm} = v_{be}g_m \qquad (3.10)$$

where $i_{C\mu}$ is the capacitor current, v_{ce} (or v_{ds}) is v_{out} (or zero), and z_{RHP} is the right-half-plane zero located at $g_m/2\pi C_\mu$ or $g_m/2\pi C_{GD}$ in hertz.

Capacitors C_μ and C_{GD} also shunt current away from the input and output, effectively decreasing their respective signal strengths and instilling shunting effects on the frequency response of the circuit. Generally, and similarly, because parasitic capacitors to ac ground exist at every node in the ac path, every node introduces a gain-shunting pole, which in this case amounts to poles at v_{in} and v_{out}. Had v_{IN} been an ideal voltage source, where its equivalent source resistance R_S is zero (Fig. 3.2), the input pole would have been at negligibly high frequencies, except R_S is not zero in practice. The fact is a fraction of the current steers away from the input through C_μ or C_{GD}, similar to the effects of a shunting capacitor to ground. What is peculiar about this capacitor is that as v_{in} increases or decreases, because the capacitor is connected across an inverting gain stage, v_{out} decreases or increases at a faster rate so that a small change in v_{in} translates to an amplified capacitor-voltage variation. The implication of this so called *Miller effect* is that a capacitor displacement current has less of an impact on v_{in} than the physical capacitor would have, had it been connected to ground. In other words, the effective input capacitance (Miller input capacitance $C_{Miller.I}$) is larger than the physical capacitance by approximately a factor equal to the gain across the capacitor:

$$Z_{C_{Miller.I}} = \frac{1}{sC_{Miller.I}} \equiv \frac{v_{in}}{i_C} = \frac{v_{in}}{(v_{in} - v_{out})sC_\mu} = \frac{v_{in}}{(v_{in} - Av_{in})sC_\mu}$$

$$= \frac{v_{in}}{v_{in}(1-A)sC_\mu} = \frac{1}{s(1-A)C_\mu} \qquad (3.11)$$

or

$$C_{\text{Miller}.I} = (1 - A)C_\mu \approx -AC_\mu \tag{3.12}$$

where $z_{\text{C.Miller}.I}$ is the effective input capacitor impedance of C_μ or C_{GD} to ground, i_C is the capacitor current, s in Laplace transforms represents frequency, and A is the inverting gain across C_μ or C_{GD} [e.g., $-g_m(r_o \,||\, R_{\text{Load}})$]. This effect, however, has minimal impact on C_μ's or C_{GD}'s effective output Miller capacitance $C_{\text{Miller.O}}$ because changes in v_{out} translate to negligibly small changes in v_{in} (since v_{in} is approximately v_{out}/A) so no capacitor-voltage amplification results and C_μ or C_{GD} only present its physical capacitance to the output:

$$\begin{aligned} Z_{C_{\text{Miller}.O}} &= \frac{1}{sC_{\text{Miller}.O}} \equiv \frac{v_{\text{out}}}{i_C} = \frac{v_{\text{out}}}{(v_{\text{out}} - v_{\text{in}})sC_\mu} \\ &= \frac{v_{\text{out}}}{\left(v_{\text{out}} - \dfrac{v_{\text{out}}}{A}\right)sC_\mu} = \frac{1}{\left(1 - \dfrac{1}{A}\right)sC_\mu} \approx \frac{1}{sC_\mu} \end{aligned} \tag{3.13}$$

or

$$C_{\text{Miller}.O} = \left(1 - \frac{1}{A}\right)C_\mu \approx C_\mu \tag{3.14}$$

The total input capacitance C_{IN} into the base or gate of a basic CE/CS circuit is the sum of Miller capacitance C_{Miller} and base-emitter or gate-source capacitances C_π or C_{GS}, the results of which are lower base-emitter or gate-source voltages v_{be} or v_{gs} at higher frequencies and therefore lower ac gain. Input capacitance C_{IN} inflicts little to no effects at lower frequencies, when its impedance is high, but when its impedance approaches $R_S \,||\, r_\pi$ (or R_S in the MOS case), the ac voltage divider starts to attenuate v_{in}, effectively introducing a pole at $1/2\pi(R_S \,||\, r_\pi)C_{\text{IN}}$ or $1/2\pi R_S C_{\text{IN}}$:

$$z_{\text{C.IN}} = \frac{1}{sC_{\text{IN}}} = \frac{1}{s[C_\pi + (1 - A)C_\mu]}\bigg|_{p_{\text{IN}} \approx \frac{1}{2\pi(-A)C_\mu(R_S||r_\pi)}} \equiv R_S \,||\, r_\pi \tag{3.15}$$

where $z_{\text{C.IN}}$ is capacitor impedance, p_{IN} is the Miller input pole of the circuit in hertz, and C_{GS} and C_{GD} replace C_π and C_μ and r_π disappears (i.e., r_π approaches infinity) in the MOS case. In applying the TC approach, p_{IN} often proves dominant, and because its dominance is primarily the result of C_μ or C_{GD} (not C_π or C_{GS}), TC analysis dictates

C_μ or C_{GD} short at lower frequencies, before C_{Load} is able to shunt energy away from v_{out}.

At frequencies higher than p_{IN}, past the point where C_μ or C_{GD} short, the impedance looking into the collector or drain of the CE/CS gain stage decreases because C_μ or C_{GD} impresses v_{out} back on the base or gate, changing the characteristics of transconductor current i_{gm}. As a result, i_{gm} is now completely dependent on v_{out} (i.e., i_{gm} is $v_{out}g_m$) and consequently a function of its own two terminals, which has the same effect as a resistance with a value of $1/g_m$:

$$i_{gm}\Big|_{C_\mu = \text{Short}} = v_{out}\, g_m \equiv \frac{v_{out}}{r_m} \tag{3.16}$$

where r_m is $1/g_m$. Because connecting the collector and base terminals amounts to using the BJT as a base-emitter pn-junction diode, the CE/CS transistor is said to be *diode connected* at high frequencies. Loading capacitor C_{Load}, at sufficiently high frequencies for its impedance to be low, shunts the parallel combination of R_{Load}, r_o or r_{ds}, r_π (which is now also present at v_{out} in the BJT case), and $1/g_m$, the latter of which is relatively low to begin with between 10 Ω and 10 kΩ. The pole that results is therefore at relatively higher frequencies, when the impedance across C_π or C_{GS}, C_{Load}, and effective output Miller capacitance $C_{Miller.O}$ or approximately C_μ or C_{DG} approximately equals $1/g_m$:

$$Z_{C.O} = \frac{1}{s(C_\pi + C_{Load})}\Bigg|_{p_O \approx \frac{g_m}{2\pi(C_\pi + C_{Load})}} \equiv R_{Load} \parallel r_o \parallel r_\pi \parallel \frac{1}{g_m} \approx \frac{1}{g_m} \tag{3.17}$$

where $Z_{C.O}$ is the capacitor impedance and p_O is the output pole located at $g_m/2\pi(C_\pi + C_{Load})$ in hertz—r_{ds} replaces r_o and r_π disappears in the MOS case.

The effects of C_π or C_{GS} and C_μ or C_{GD} on the degenerated CE or CS stage are similar, introducing what amounts to a low-frequency input Miller pole p_{IN}, a high-frequency output pole p_O, and a right-half-plane zero z_{RHP}. Degenerating resistor R increases the input resistance of the CE case and decreases the inverting gain of the amplifier (i.e., reduces the Miller effect). Of these two effects, the latter may have more impact on p_{IN} than the former if source resistor R_S were to be lower than the degenerated input resistance, pushing p_{IN} to higher frequencies (now that C_{IN} is also low). At higher frequencies, in the output, neglecting the effects of r_o and r_{ds} because they are substantially higher than $1/g_m$, R is in series with the now $1/g_m$ resistor, pulling p_O to lower frequencies. Because the effective transconductance degenerates with R, R also pulls z_{RHP} to lower frequencies. In practice, however, degenerating resistor R is normally low relative to r_π, r_o, and r_{ds} and its effects on frequency response are not as pronounced.

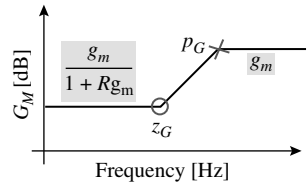


FIGURE 3.6 Frequency response of the degenerated CE/CS transistor's transconductance G_M in the presence of a degenerating capacitance.

The parasitic capacitance (C_{PAR}) present across degenerating resistor R can also affect the circuit at frequencies of interest through degenerated transconductance G_M, but only if the impedance it presents is sufficiently low with respect to R (i.e., C_{PAR} is large) to steer current away from R. In qualitative terms, G_M increases when C_{PAR} starts to shunt degenerating resistance R, as shown in Fig. 3.6, past pole p_G:

$$Z_{C.PAR} = \left.\frac{1}{sC_{PAR}}\right|_{z_G \approx \frac{1}{2\pi RC_{PAR}}} \equiv R \tag{3.18}$$

Transconductance G_M, however, ultimately reaches its nondegenerated state of g_m (and flattens with respect to frequency) when parasitic impedance $1/sC_{PAR}$ is negligibly small with respect to R, past pole p_G. The location of p_G corresponds to how much frequency must traverse to increase low-frequency degenerated transconductance $G_{M.LF}$ [or roughly $g_m/(1 + g_m R)$] to its nondegenerated target of g_m (or roughly g_m), and because gain increases linearly past zero z_G, p_G is $g_m/G_{M.LF}$ times higher than z_G:

$$G_M = G_{M.LF}\left(1+\frac{2\pi s}{z_G}\right)\Bigg|_{f>>z_G} \approx G_{M.LF}\left(\frac{2\pi s}{z_G}\right)\Bigg|_{p_G \approx \frac{g_m z_G}{G_{M.LF}}} \equiv g_m \tag{3.19}$$

Replacing degenerating resistance R with the parallel combination of C_{PAR}'s impedance and R also reveals the location of the aforementioned zero-pole pair:

$$\begin{aligned} G_M &\approx \frac{g_m}{1+g_m\left(R \parallel \frac{1}{sC_{PAR}}\right)} = \frac{g_m}{1+g_m\left(\frac{R}{1+sRC_{PAR}}\right)} \\ &= \frac{g_m(1+sRC_{PAR})}{(1+g_mR)\left[1+\frac{sRC_{PAR}}{(1+g_mR)}\right]} \frac{G_{M.LF}\left(1+\frac{2\pi s}{z_G}\right)}{\left[1+\frac{2\pi s}{z_G}\left(\frac{G_{M.LF}}{g_m}\right)\right]} \equiv \frac{G_{M.LF}\left(1+\frac{2\pi s}{z_G}\right)}{\left(1+\frac{2\pi s}{p_G}\right)} \end{aligned} \tag{3.20}$$

Note this zero-pole effect is more prevalent in MOS devices when bulk and source terminals share a common node because the parasitic bulk to substrate capacitance in welled MOS devices can be substantial.

3.1.2 Common-Collector/Drain, Emitter/Source Voltage Followers

Common-collector (CC) and common-drain (CD) circuits have their collectors and drains tied to low-impedance points (i.e., ac grounds), as shown in Fig. 3.7*a*. Since collector and drain currents only flow through collector, emitter, drain, and source terminals, bases and gates are poor outputs. As a result, emitters and sources, the only plausible output terminals in CC/CD configurations, are outputs. Similarly, because collector and drain currents are strong functions of base and gate voltages, and because they are the only other terminals available, bases and gates are inputs.

Large-Signal Operation

Unlike their CE and CS counterparts, their output voltages are low when their input voltages are low, and vice versa, which means they are noninverting in nature. The input voltage at which the output starts to rise ($V_{IN(low)}$) in the n-type CC or CD configuration shown corresponds to the dc base-emitter or gate-source voltage V_{BE} or V_{GS} necessary to sustain load current I_{Bias}, where V_{BE} is approximately 0.65–0.8 V and V_{GS} the sum of $|V_T|$ and $V_{DS(sat)}$, which can be 1–1.5 V. Because of the same reason, the maximum output voltage of the circuit is below the supply by a V_{BE} or V_{GS} (e.g., 0.65–1.5 V). Minimum output voltage $V_{OUT(min)}$, on the other hand, corresponds to the CC or CD transistor being off, at which point no current flows and the biasing

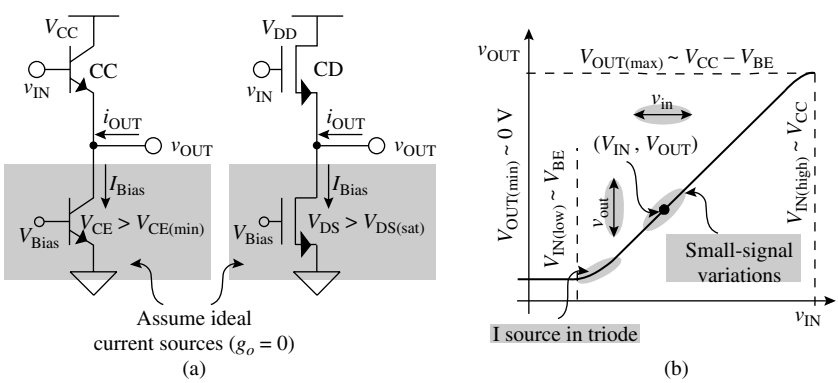


Figure 3.7 (*a*) Common-collector and common-drain emitter/source followers and (*b*) their representative large-signal response.

transistor, while entering the triode region, pulls v_{OUT} to its supply (e.g., $V_{OUT(min)}$ is zero in the case shown).

Small-Signal Response at Low Frequency

As in the CE/CS case, an analog designer usually biases the CC or CD transistor in its high-gain mode (i.e., forward active or slightly saturated for the BJT and saturation for the MOSFET), where the circuit has the highest gain. In this case, however, the range of V_{IN} and V_{OUT} for which the gain is relatively high and consistent is wide, as can be appreciated from the large-signal response shown in Fig. 3.7*b*. The driving reason for this wide linear range is that the gain is close to unity, since v_{BE} and v_{GS} remain unchanged at whatever voltage is necessary to sustain bias current I_{Bias} throughout most of the v_{IN} range. This unchanging voltage, for all practical purposes, amounts to a dc offset voltage between base and emitter or gate and source, in other words, a battery with no signal-gain characteristics between v_{IN} and v_{OUT}, which is why these circuits are more commonly known as *emitter* and *source followers*.

Replacing the CC and CD transistors with their respective small-signal equivalent models, as shown in Fig. 3.8, provides a more accurate means of predicting the voltage gain of the circuit. Before starting, though, it is helpful to realize the current flowing through the bulk-effect transconductor is proportional to its own two terminal voltages so a $1/g_{mb}$ resistor can represent its effect:

$$R_{gmb} = \frac{v_{out}}{i_{gmb}} = \frac{v_{out}}{v_{out} g_{mb}} = \frac{1}{g_{mb}} \tag{3.21}$$

where R_{gmb} is the equivalent small-signal resistance into transconductor g_{mb}. Ultimately, the output voltage is the ohmic voltage drop across the

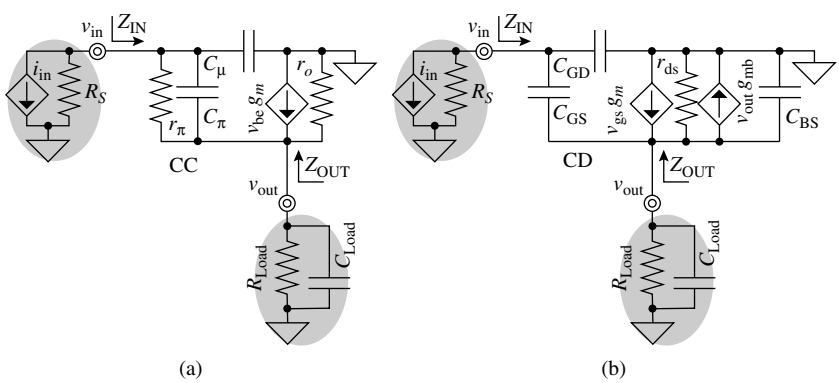


Figure 3.8 Small-signal models of (*a*) emitter and (*b*) source followers with their respective source- and load-equivalent circuits.

parallel combination of output resistor r_o or r_{ds}, $1/g_{mb}$, and R_{Load} caused by transconductor current $v_{be}g_m$ or $v_{gs}g_m$, resulting in a low-frequency voltage gain A_{V0} that is slightly below unity:

$$A_{V0} \equiv \frac{v_{out}}{v_{in}} = \frac{v_{out}}{v_{gs} + v_{out}} = \frac{v_{gs}g_m\left(r_{ds} \,||\, R_{Load} \,||\, \frac{1}{g_{mb}}\right)}{v_{gs} + v_{gs}g_m\left(r_{ds} \,||\, R_{Load} \,||\, \frac{1}{g_{mb}}\right)}$$

$$= \frac{g_m\left(r_{ds} \,||\, R_{Load} \,||\, \frac{1}{g_{mb}}\right)}{1 + g_m\left(r_{ds} \,||\, R_{Load} \,||\, \frac{1}{g_{mb}}\right)} \approx 0.7 - 0.9 \qquad (3.22)$$

where the gain corresponds to the MOS source follower. In the case of the BJT, bulk-effect resistor $1/g_{mb}$ disappears, v_{be} and r_o replace v_{gs} and r_{ds}, and an additional base current equal to v_{be}/r_π flows through $r_{ds} \,||\, R_{Load}$, but because r_π is β times larger than $1/g_m$, the gain relationship is similar:

$$A_{V0} \equiv \frac{v_{out}}{v_{in}} = \frac{v_{out}}{v_{be} + v_{out}} = \frac{v_{be}\left(g_m + \frac{1}{r_\pi}\right)(r_o \,||\, R_{Load})}{v_{be} + v_{be}\left(g_m + \frac{1}{r_\pi}\right)(r_o \,||\, R_{Load})}$$

$$\approx \frac{g_m(r_o \,||\, R_{Load})}{1 + g_m(r_o \,||\, R_{Load})} \approx 0.9 - 0.99 \qquad (3.23)$$

Because the bulk effect is absent and the transconductance of BJTs is higher than MOSFETs, given i_C's exponential relationship to v_{BE}, the emitter follower gain is closer to unity than the MOS counterpart.

The small-signal circuit of the emitter follower for the analysis of its input resistance is the same as that of the emitter-degenerated CE case, except the series degenerating resistor is now the parallel combination of R_{Load} and r_o whose collective effect increases the input resistance of the circuit:

$$R_{IN} = \frac{v_{in}}{i_{in}} = \frac{v_\pi + v_R}{i_b} = \frac{i_b r_\pi + (1+\beta) i_b (R_{Load} \| r_o)}{i_b} = r_\pi + (1+\beta)(R_{Load} \| r_o) \qquad (3.24)$$

where the total current flowing through $R_{Load} \,||\, r_o$ is base current i_b and collector current i_c or βi_b. Again, as in the CS case, the effective degenerating resistor has no effect on R_{IN} in the CD circuit because the resistance into a gate is already infinitely large.

The output resistance of the CC emitter-follower circuit (R_{OUT}) is the parallel combination of R_{Load}, r_o, the series combination of r_π and source resistance R_S (since i_{in} is zero when determining R_{OUT}), and the equivalent resistance into the transconductor (R_{gm}), the latter of which is a base-degenerated version of $1/g_m$ (since R_S degenerates v_{be}):

$$R_{gm} = \frac{v_e}{i_{gm}} = \frac{v_e}{-v_{be}g_m} = \frac{v_e}{\left(\frac{v_e r_\pi}{r_\pi + R_S}\right)g_m} = \frac{r_\pi + R_S}{r_\pi g_m} = \frac{r_\pi + R_S}{\beta} \geq \frac{1}{g_m} \qquad (3.25)$$

R_{gm} reduces to $1/g_m$ if R_S is moderately low, which may not be the case. All the same, R_{gm} remains relatively low because β is large, so R_{OUT} simplifies to R_{gm}, as R_{gm}'s impedance is for the most part considerably lower than the other parallel resistances:

$$R_{OUT} = R_{Load} \parallel r_o \parallel (r_\pi + R_S) \parallel R_{gm} \approx R_{gm} \approx \frac{r_\pi + R_S}{r_\pi g_m} \geq \frac{1}{g_m} \qquad (3.26)$$

Output resistance R_{OUT} for the MOS case is similar, except the series combination of r_π and R_S and their degenerating effect on $1/g_m$ disappear while bulk effects appear as another parallel resistance (i.e., $1/g_{mb}$):

$$r_{OUT} = R_{Load} \parallel r_o \parallel \frac{1}{g_{mb}} \parallel \frac{1}{g_m} \approx \frac{1}{g_m} \qquad (3.27)$$

where g_{mb} is considerably smaller than g_m by a factor of 10 or so. In spite of the seemingly lower R_{OUT} for the MOS case, given r_π and R_S no longer have a degenerating effect on $1/g_m$, R_{OUT} is often larger than its BJT counterpart because its g_m is normally smaller than the BJT's. Note the low output impedance of the follower circuit, in general, is one of its most appealing features.

Small-Signal Response at High Frequency

Because every ac node has parasitic capacitance to ground, incoming signals lose energy to what amounts to poles at the base (or gate) and emitter (or source). What is perhaps not as apparent at first glance is that C_π or C_{GS} feed-forward an in-phase signal from v_{in} to v_{out}, introducing a left-half plane zero to the mix. As before, the feed-forward path implies a zero and, because it is in phase with r_π and g_m currents and helps r_π and g_m drive v_{out}, the zero is relatively benign and on the left-half plane. The effects of the zero are nonetheless felt when the current through C_π or C_{GS} (i.e., i_C) supersedes the combined

contribution of g_m (i.e., i_{gm}) and r_π (or base current i_b), when i_C equals and exceeds the sum of i_b and i_{gm}:

$$i_C = v_{be} sC_\pi \Big|_{z_{LHP} \approx \frac{g_m}{2\pi C_\pi}} \equiv i_{gm} + i_b = v_{be}\left(g_m + \frac{1}{r_\pi}\right) \approx v_{be} g_m \tag{3.28}$$

where left-half-plane zero z_{LHP} reduces to $g_m/2\pi C_\pi$ because base current i_b is β times smaller than $v_{be}g_m$. In the MOS case, v_{gs} and C_{GS} replace v_{be} and C_π, C_{BS} adds to C_{Load} and the effects of r_π disappear (i.e., r_π is infinite), reducing z_{LHP} to $g_m/2\pi C_{GS}$.

The pole associated with the input (i.e., base or gate) occurs when the effective parasitic capacitance at that node steers (or shunts) current away from the resistance present, and C_π or C_{GS}, C_μ or C_{GD}, and C_{Load} or $C_{Load} + C_{BS}$ all have an impact in this. In fact, similar to the increasing effect R_{Load} has in R_{IN} of the emitter follower (i.e., R_{IN} is r_π in series with $(1 + \beta)R_{Load}$), impedance $1/sC_{Load}$ increments $1/sC_\pi$ or $1/sC_{GS}$ by $(1 + \beta)/sC_{Load}$,

$$\begin{aligned} z_C \equiv \frac{1}{sC_{IN.BE}} &= \frac{v_{in.c}}{i_{in.c}} = \frac{v_{\pi.c} + v_{out.c}}{i_{in.c}} = \frac{\dfrac{i_{in.c}}{sC_\pi} + \dfrac{(1+\beta)i_{in.c}}{sC_{Load}}}{i_{in.c}} \\ &= \frac{1}{sC_\pi} + \frac{(1+\beta)}{sC_{Load}} \approx \frac{\beta}{sC_{Load}} \end{aligned} \tag{3.29}$$

which means the input capacitance associated with the base-emitter path ($C_{IN.BE}$) is the series combination of C_π and $C_{Load}/(1 + \beta)$:

$$C_{IN.BE} = C_\pi \oplus \frac{C_{Load}}{1+\beta} \approx C_\pi \oplus \frac{C_{Load}}{\beta} \tag{3.30}$$

where z_C is the capacitor impedance and lower case c" in the subscripts implies only capacitive effects are taken into account. Considering β is on the order of 100, $C_{Load}/(1 + \beta)$ is often (but not always) smaller than C_π and $C_{IN.BE}$ therefore reduces to C_{Load}/β. As a result, C_μ or C_{GD} and $C_{IN.BE}$ together attenuate v_{in} when their collective impedance $1/sC_{IN}$ nears the resistance at v_{in}, the parallel combination of R_S and R_{IN}:

$$\begin{aligned} z_C = \frac{1}{sC_{IN}} &\approx \frac{1}{s\left[C_\mu + \left(C_\pi \oplus \dfrac{C_{Load}}{1+\beta}\right)\right]}\Bigg|_{p_{IN} \approx \frac{1}{2\pi(R_S \| R_{IN})\left[C_\mu + \left(C_\pi \oplus \frac{C_{Load}}{\beta}\right)\right]}} \\ &\equiv R_S \| R_{IN} = R_S \| [r_\pi + (1+\beta)(R_{Load} \| r_o)] \end{aligned} \tag{3.31}$$

When source resistance R_S is large, as when driven by a transconductor, the pole is at relatively low frequencies, and vice versa, when driven by a voltage source, the latter of which is less often the case in ICs.

In the MOS version of the circuit, similar effects occur, except R_{IN} is infinitely large, C_{GD} replaces C_μ, C_{GS} replaces C_π, C_{BS} adds to C_{Load}, and effective current gain β is infinite at dc and decreases with decreasing C_{GS} and C_{GD} impedances (as frequency increases):

$$\beta_{MOS} \equiv \frac{i_d}{i_g} \approx \frac{v_{gs} g_m}{i_g} = \frac{\left(\dfrac{i_g}{s(C_{GS}+C_{GD})}\right) g_m}{i_g} = \frac{g_m}{s(C_{GS}+C_{GD})} \qquad (3.32)$$

Normally, the current gain attenuates the effects of C_{Load} and C_{BS} to the point where C_μ or C_{GD} is considerably larger than the $1/\beta$ or $1/\beta_{MOS}$ translation of C_{Load} and C_{BS} and p_{IN} reduces to $1/2\pi(R_S \,||\, R_{IN})C_\mu$ or $1/2\pi R_S C_{GD}$, neither of which is at negligibly high frequencies with moderate-to-high source resistances. Note C_π or C_{GS} decreases impedance Z_{IN} with frequency, giving Z_{IN} a capacitive quality.

Finding the location of output pole p_O is not always straightforward because source resistance R_S in emitter followers may be large enough to increase R_{OUT} considerably beyond normal $1/g_m$ levels, somewhat obscuring the conclusion that p_O must, by default, reside at high frequencies. Additionally, as C_π or C_{GS} short and small-signal base-emitter or gate-source voltage v_{be} or v_{gs} decrease with increasing frequencies, transconductance current i_{gm} and consequently the voltage gain A_V across the stage decrease. This decrease in i_{gm}, incidentally, also amounts to an inductive effect on R_{OUT} because it causes output impedance Z_{OUT} to increase with frequency. In any event, when R_S is low, R_{OUT} is also low and p_O is at high frequencies, and in the case R_S is high, input pole p_{IN} precedes p_O and the equivalent source impedance at the input is still low at higher frequencies, which means, irrespective of R_S, Z_S is low by the time p_O asserts its influence.

With respect to p_O, equating source impedance Z_S to zero and applying the TC method reveals, when referring to the OCTC equivalent circuit shown in Fig. 3.9, C_π or C_{GS} and C_{Load} or $C_{Load} + C_{BS}$ together shunt ac signal energy at v_{out} to ac ground. To determine the equivalent resistance these capacitors short, as a visual aid, the figure decomposes transconductance component $v_{be}g_m$ into its constituent parts: $v_b g_m$ and $-v_e g_m$, which reduce, respectively, to $(0)g_m$ and $v_{out}g_m$ with the latter directing its current to ground, like a $1/g_m$ resistor. The equivalent parallel resistance to C_π or C_{GS} and C_{Load} or $C_{Load} + C_{BS}$ is therefore the parallel combination of r_π (in the case of a BJT), $1/g_m$, r_o or r_{ds}, R_{Load},

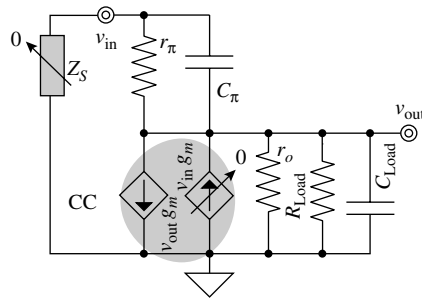


FIGURE 3.9
Time-constant (TC) equivalent circuit for ascertaining output pole p_o (the time constant associated with C_π and C_{Load} in the CC emitter follower circuit).

and in the case of a MOSFET, $1/g_{mb}$, introducing a pole at approximately $g_m/2\pi(C_\pi + C_{Load})$:

$$z_C = \frac{1}{s(C_\pi + C_{Load})}\bigg|_{p_O \approx \frac{g_m}{2\pi(C_\pi + C_{Load})}} \equiv r_\pi \,\|\, \frac{1}{g_m} \,\|\, r_o \,\|\, R_{Load} \tag{3.33}$$

where z_C is the impedance across the capacitors. From an algebraic standpoint, with respect to A_V, while C_π or C_{GS} alters the forward transconductance of the follower (alongside g_m and $1/r_\pi$ or simply g_m), C_{Load} affects the load (with r_o and R_{Load}):

$$A_V \equiv \frac{v_{out}}{v_{in}} = \frac{v_{out}}{v_{be} + v_{out}} = \frac{v_{be}\left(g_m + sC_\pi + \frac{1}{r_\pi}\right)\left(r_o \,\|\, R_{Load} \,\|\, \frac{1}{sC_{Load}}\right)}{v_{be} + v_{be}\left(g_m + sC_\pi + \frac{1}{r_\pi}\right)\left(r_o \,\|\, R_{Load} \,\|\, \frac{1}{sC_{Load}}\right)}$$

$$\approx \frac{(g_m + sC_\pi)\left(r_o \,\|\, R_{Load} \,\|\, \frac{1}{sC_{Load}}\right)}{1 + (g_m + sC_\pi)\left(r_o \,\|\, R_{Load} \,\|\, \frac{1}{sC_{Load}}\right)} = \frac{(g_m + sC_\pi)}{\left[\frac{1 + sC_{Load}(r_o \,\|\, R_{Load})}{r_o \,\|\, R_{Load}}\right] + (g_m + sC_\pi)}$$

$$= \frac{(g_m + sC_\pi)}{\left(\frac{1}{r_o \,\|\, R_{Load}} + g_m\right) + sC_{Load} + sC_\pi} = \frac{g_m\left(r_o \,\|\, R_{Load} \,\|\, \frac{1}{g_m}\right)\left(1 + \frac{sC_\pi}{g_m}\right)}{\left[1 + s(C_\pi + C_{Load})\left(r_o \,\|\, R_{Load} \,\|\, \frac{1}{g_m}\right)\right]}$$

$$= \frac{A_{v0}\left(1 + \frac{sC_\pi}{g_m}\right)}{\left[1 + s(C_\pi + C_{Load})\left(r_o \,||\, R_{Load} \,||\, \frac{1}{g_m}\right)\right]} \approx \frac{A_{v0}\left(1 + \frac{sC_\pi}{g_m}\right)}{\left[1 + \frac{s(C_\pi + C_{Load})}{g_m}\right]} \tag{3.34}$$

which shows p_O is independent of R_S and a function of both C_π (or C_{GS}) and C_{Load} (or C_{Load} and C_{BS}), falling at $g_m/2\pi(C_\pi + C_{Load})$ or $g_m/2\pi(C_{GS} + C_{Load} + C_{BS})$.

Note output pole p_O precedes zero z_{LHP} because the capacitance associated with the former is incremented by C_{Load} while the resistance is the same ($1/g_m$). Nevertheless, when lightly loaded, the pole and zero may not be far from one another, which means they tend to cancel and collectively produce little to no noticeable effects on the frequency response of the circuit. Even if C_{Load} is higher, though, p_O remains at relatively higher frequencies because Z_{OUT} is relatively low, which as before, but now within the context of frequency, is one of the most important features of this circuit: wide bandwidth.

3.1.3 Common-Base/Gate Current Buffers

The common-base (CB) and common-gate (CG) circuits have their bases and gates tied to ac ground, as shown in Fig. 3.10*a*. Because collector and drain currents are strong functions of base-emitter and gate-source voltages, respectively, and weak functions of collector and drain voltages, bases, emitters, gates, and sources are good inputs while collectors and drains are not. As such, the only other input available in a CB or CG circuit is its emitter or source. Similarly, because gain current flows through collectors, emitters, drains, and sources and emitters and sources are already assigned in CB and CG circuits, collectors and drains are the only outputs available.

Although voltage-gain characteristics may describe the behavior of the CB/CG circuit, its current-buffering features are apparent because any current pulled or pushed into the emitter or source ends up flowing through the collector or drain (or most of it, in the case of the BJT, as a minuscule portion is lost as base current). In any case,

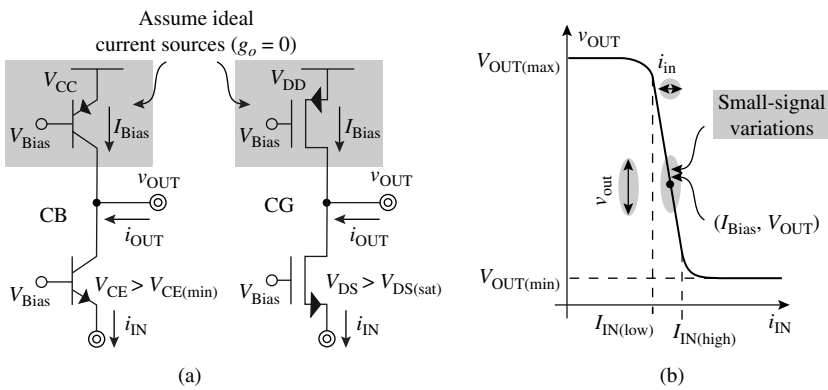


Figure 3.10 (*a*) Common-base and common-gate current-buffer circuits and (*b*) their large-signal response.

applying a voltage to the emitter or source sets the base-emitter or gate-source voltage of the CB/CG transistor, given a bias base or gate voltage, and defines, as a result, its effective input current i_{IN}. In practice, however, since emitters and sources present low resistances, as seen from the output resistances of CC/CD emitter/source followers, emitters and sources are good input current points, and because emitter and source currents roughly equal collector and drain currents, respectively, CB and CG transistors are also good *current buffers*.

Large-Signal Operation

In the case of a resistive load, pulling more current from the input terminal of a CB/CG current buffer pulls current through its load and decreases its terminal voltage, and vice versa. The response for a current-source load is no different, except the loading resistance is larger and its dc current bias (I_{Bias}) offsets its midbias point, as shown in the large-signal response of Fig. 3.10*b*. In other words, when the dc portion of pulling current i_{IN} (i.e., I_{IN}) is greater than I_{Bias}, I_{IN} overwhelms I_{Bias} and pulls dc output voltage V_{OUT} low to within a minimum BJT collector-emitter (0.2–0.3 V) or triode MOS drain-source ($I_{IN}R_{Triode}$) voltage from emitter/source voltage V_{IN}, which is below base or gate bias voltage V_{Bias} by only one V_{BE} or V_{GS}. On the other hand, if I_{IN} is less than I_{Bias}, I_{Bias} overwhelms I_{IN} and pulls V_{OUT} to within a minimum BJT collector-emitter or triode MOS drain-source voltage from the positive supply. Only if I_{IN} equals I_{Bias} does the circuit balance and V_{OUT} is able to reside between the supply rails.

Small-Signal Response at Low Frequency

The high-gain mode for this circuit is when the CB/CG transistor and its load are slightly saturated or in the forward-active region (or in saturation for the case of the MOSFET), which occurs when the dc portion of input current i_{IN} equals biasing current I_{Bias}, as shown in Fig. 3.10*b*. At this point, small changes in i_{IN} (i.e., i_{in}) cause relatively large ac variations in v_{OUT} (i.e., v_{out}). Note that pushing more current into the emitter or source, which amounts to decreasing i_{IN} as defined in the n-type case shown in Fig. 3.10*a*, causes an increase in output voltage, which is why the CB/CG circuit is said to be a noninverting circuit, much like the CC/CD emitter/source follower and unlike the CE/CS amplifier.

To ascertain its small-signal characteristics, it is useful to replace the CB/CG transistor with its small-signal model, as illustrated in Fig. 3.11. To start, input resistance R_{IN} is the parallel combination of r_π and the equivalent resistance seen into the transconductor and output resistance r_o (i.e., R_{EQ}):

$$R_{IN} = r_\pi \parallel R_{EQ} \tag{3.35}$$

where r_π disappears and r_{ds} replaces r_o in R_{EQ} in the MOS case. Note, however, referring to Fig. 3.10*a*, if v_{out} were ac-grounded, R_{IN} would

mimic the output resistance of the CC/CD follower and R_{EQ} would be the parallel combination of r_π, $1/g_m$, and r_o, whose total resistance is relatively low at $1/g_m$. In the case of the CB/CG stage shown, however, there is a series resistor (i.e., R_{Load}) between the collector/drain and ac ground, and if this resistance were infinitely large, so would R_{EQ}, given that it is in series and there is no other current path to ac ground available for R_{EQ} current. As a result, considering input voltage v_{in} is

$$v_{in} = (i_{in} - v_{in} g_m) r_o + i_{in} R_{Load} \tag{3.36}$$

R_{Load} loads the transconductance current and equivalent resistance R_{EQ} is a loaded version of $1/g_m$:

$$R_{EQ} \equiv \frac{v_{in}}{i_{in}} = \frac{\left(\frac{i_{in} r_o + i_{in} R_{Load}}{1 + g_m r_o} \right)}{i_{in}} = \frac{r_o + R_{Load}}{1 + g_m r_o} \geq \frac{1}{g_m} \tag{3.37}$$

where r_{ds} replaces r_o and the sum of g_m and bulk-effect g_{mb} (i.e., $g_m + g_{mb}$) replace g_m in the MOS case, the latter of which results because g_m and g_{mb} are in parallel and under the control of the same variable (i.e., v_{in}). Note that using a loading resistor equivalent to r_o or r_{ds}, which is reasonable, considering it represents a nondegenerated CE/CS transistor, reduces R_{EQ} and consequently R_{IN} to approximately $2/g_m$,

$$R_{EQ} = \left. \frac{r_o + R_{Load}}{1 + g_m r_o} \right|_{R_{Load} = r_o} \approx \frac{2}{g_m} \tag{3.38}$$

which is still a relatively low value. A degenerated CE/CS transistor load may on the other hand increase R_{EQ} to larger values, to r_π or r_{ds}, for instance, if the degeneration resistor is another CE/CS device:

$$R_{EQ} = \left. \frac{r_o + R_{Load}}{1 + g_m r_o} \right|_{R_{Load} = g_m r_o (r_\pi \| r_o)} \approx \frac{r_o + g_m r_o (r_\pi \| r_o)}{1 + g_m r_o} \approx \frac{g_m r_o r_\pi}{g_m r_o} \approx r_\pi \tag{3.39}$$

The output resistance of the CB/CG current buffer is similar to the degenerated CE/CS amplifier, being that both outputs are collectors and both circuits have a degenerating resistor. In the CE/CS case, v_{in} is ac-grounded to null out the effects of the input on output resistance R_{OUT} because v_{in}'s effects are already accounted in transconductance G_M. In the CB/CG case, i_{in} is open-circuited to null out its effects on R_{OUT}, but source resistance R_S remains. As a result, in determining R_{OUT} and consequently open-circuiting i_{in} and ac-grounding v_{in} in the CB/CG and degenerated CE/CS small-signal equivalent circuits of Figs. 3.3*b* and 3.11, respectively, their resulting circuits equal

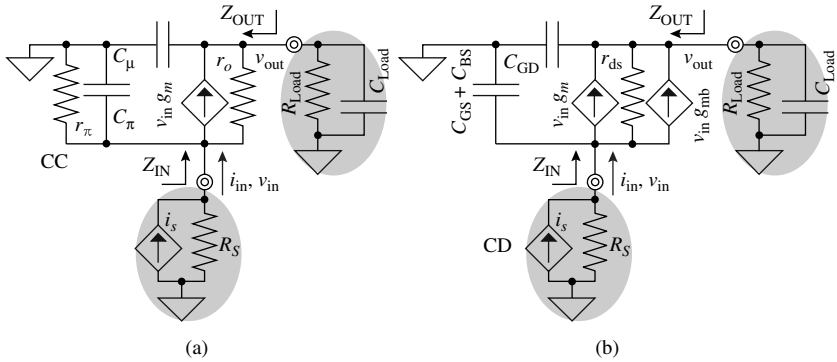


FIGURE 3.11 Small-signal models of the (*a*) common-base (CB) and (*b*) common-gate (CG) current buffers.

(i.e., degenerating resistor R in Fig. 3.3*b* is source resistance R_S in Fig. 3.11) and so do their R_{OUT}'s:

$$\begin{aligned} R_{OUT}\Big|_{i_{in}=0} &= \frac{v_{out}}{i_{out}} = \frac{(i_{out} - v_{gs}g_{m.eff})r_{ds} - v_{gs}}{i_{out}} \\ &= \frac{(i_{out} + R_S i_{out} g_{m.eff})r_{ds} + R_S i_{out}}{i_{out}} = r_{ds} + R_S(g_m + g_{mb})r_{ds} + R_S \\ &\approx R_S(g_m + g_{mb})r_{ds} \end{aligned} \tag{3.40}$$

where v_{be}, g_m, r_o, and r_π in parallel with R_S (i.e., $r_\pi \| R_S$) replace v_{gs}, $g_{m.eff}$ (i.e., $g_m + g_{mb}$), r_{ds}, and R_S, respectively, in the BJT case.

The gain of the current buffer takes one of three forms: current, transresistance, or voltage gain. As a current amplifier, input current i_{in} splits between R_S and the CB/CG transistor, and in the CB case, r_π further steers current away from the output. As a result, the current gain through the CB/CG transistor is the ratio of the fraction of current that reaches the collector/drain terminal and the current driven into the circuit. In other words, input current i_{in} induces an ohmic voltage drop across the parallel combination of input resistance R_{IN} and R_S whose ultimate effect is to vary v_{be} or v_{gs} and therefore incur a variation in current that is equivalent to a g_m translation:

$$\begin{aligned} A_{I0} \equiv \frac{i_{out}}{i_{in}} &= \left(\frac{v_{in}}{i_{in}}\right)\left(\frac{i_{out}}{v_{in}}\right) \approx \left[\frac{i_{in}(R_S \| R_{IN})}{i_{in}}\right]\left(\frac{v_{in}g_m}{v_{in}}\right) \\ &= (R_S \| R_{IN})g_m\Big|_{R_{IN} \approx \frac{1}{g_m} << R_S} \approx 1 \end{aligned} \tag{3.41}$$

which is approximately one when R_{IN} is approximately $1/g_m$ and considerably smaller than equivalent source resistance R_S. Its output voltage v_{out} is the ohmic drop across loading resistor R_{Load} caused by i_{out} (i.e., fraction of i_{in} that reaches the output) flowing through R_{Load} so its transresistance gain A_{R0} is the product of the loading resistance of the circuit and the fraction of input current that reaches the output:

$$A_{R0} \equiv \frac{v_{out}}{i_{in}} = \left(\frac{i_{out}}{i_{in}}\right)\left(\frac{v_{out}}{i_{out}}\right) \approx [(R_S \,||\, R_{IN})g_m]\left(\frac{i_{out}R_{Load} \,||\, R_{OUT}}{i_{out}}\right)$$

$$= (R_S \,||\, R_{IN})g_m(R_{Load} \,||\, R_{OUT})\Big|_{R_{IN} \approx \frac{1}{g_m} << R_S} \approx R_{Load} \,||\, R_{OUT} \qquad (3.42)$$

Although less appropriate for the circuit, given its relatively low input resistance, the voltage gain of the circuit, when considering a Thevenin equivalent input source v_{in}' (i.e., v_{in}' is $i_{in}R_S$), is the product of the voltage gain from input source v_{in}' to the emitter/source terminal of the circuit (i.e., v_{in}), which is a voltage divider, and the gain from that point to i_{out} and then to v_{out}:

$$A_{V0} \equiv \frac{v_{out}}{v_{in}'} = \left(\frac{v_{in}}{v_{in}'}\right)\left(\frac{i_{out}}{v_{in}}\right)\left(\frac{v_{out}}{i_{out}}\right) \approx \left(\frac{R_{IN}}{R_S + R_{IN}}\right)\left[\frac{v_{in}g_m}{v_{in}}\right](R_{Load} \,||\, R_{OUT})$$

$$= \left(\frac{R_{IN}}{R_S + R_{IN}}\right)g_m(R_{Load} \,||\, R_{OUT}) \qquad (3.43)$$

where r_{ds} replaces r_o and $g_m + g_{mb}$ replaces g_m in the MOS case.

Small-Signal Response at High Frequency

Capacitor C_π or C_{GS} introduces a signal-shunting pole, as it steers current away from the input of the CB/CG current buffer at higher frequencies. The resulting input pole p_{IN} occurs when the capacitor impedance is near the total resistance at the input, which is the parallel combination of R_S and R_{IN}:

$$Z_C = \frac{1}{sC_\pi}\Big|_{p_{IN} = \frac{1}{2\pi C_\pi (R_S || R_{IN})} \approx \frac{g_m}{2\pi C_\pi}} \equiv R_S \,||\, R_{IN} = R_S \,||\, r_\pi \,||\left(\frac{r_o + R_{Load}}{1 + g_m r_o}\right) \approx \frac{1}{g_m} \qquad (3.44)$$

where Z_C is capacitor impedance, $R_S \,||\, R_{IN}$ is normally low, p_{IN} is typically at high frequencies, and $C_{GS} + C_{BS}$ and r_{ds} replace C_π and r_o and r_π disappears in the MOS case. Note, however, that as C_π or $C_{GS} + C_{BS}$ short-circuit at higher frequencies, transconductor current i_{gm} decreases (i.e., v_{be} or v_{gs} decreases) and R_{IN} consequently increases, introducing an inductive effect: impedance increases with increasing frequencies. Source resistance R_S, however, limits this effect because it eventually dominates the total impedance at the input (when $R_S \,||\, R_{IN} \approx R_S$).

Although three capacitors exist in the circuit (Fig. 3.11), C_μ or C_{GD} and C_{Load} are in parallel and therefore conform to one. The effect of this collective capacitance is to steer current away from loading resistor R_{Load} and consequently decrease output voltage v_{out}. The resulting output pole p_O occurs when the impedance across C_{Load} and C_μ or C_{GD} equals the total resistance connected across their respective terminals, which amounts to the parallel combination of R_{Load} and R_{OUT}:

$$Z_C = \frac{1}{s(C_{GD}+C_{Load})}\bigg|_{p_O=\frac{1}{2\pi(C_{GD}+C_{Load})(R_{Load}\|R_{OUT})}\approx\frac{1}{2\pi(C_{GD}+C_{Load})R_{Load}}}$$

$$\equiv R_{Load}\|R_{OUT} \approx R_{Load}\|[R_s(g_m+g_{mb})r_{ds}] \approx R_{Load} \tag{3.45}$$

where Z_C is capacitor impedance and C_μ, g_m, r_π (in parallel with R_S), and r_o replace C_{GD}, $g_m + g_{mb}$, R_S, and r_{ds}, respectively, in the BJT case. Because R_{OUT} is relatively large, R_{Load} often dominates, but not necessarily.

3.1.4 Summary

Table 3.1 summarizes the different characteristics of the three single-transistor amplifiers shown and discussed. In general, the CE/CS circuit is mostly used as a voltage or transconductance amplifier, the CC/CD follower as a low output-impedance voltage follower, and the CB/CG buffer as a current buffer or transimpedance amplifier. The CE/CS configuration yields the highest output voltage swing and the CB/CG the lowest, as determined by the base/gate bias

	CE/CS V Amplifier	CC/CD V Follower	CB/CG I Buffer
Main Use	Voltage gain Transconductance gain	Voltage buffer	Transimpedance gain Current buffer
V_{OUT} Swing	High	Moderate	Low
Gain	$A_{v0} \approx -G_M(R_{OUT}\|\|R_{Load})$	$A_{v0} \approx$ +0.75–1 V/V	$A_{i0} \approx 1$ A/A, $A_{R0} \approx +R_{Load}$
R_{OUT}	High	Low	High
R_{IN}	Moderate-high	High	Low
Output Pole p_0	Low-high	High	Low
Input Pole p_{IN}	Low-moderate	Low	High
Peculiarities	Right-half-plane zero	Pole/zero pair Z_{OUT} is inductive Z_{IN} is capacitive	Z_{IN} is inductive

Table 3.1 Qualitative Comparison of Single-Transistor Circuits

voltage of the CB/CG transistor. The CE/CS circuit produces a high, but inverting voltage gain, whereas the follower and current buffer are noninverting in nature. Ultimately, irrespective of connectivity, the emitter and source terminals tend to have the lowest resistances, defining, as a result, high-frequency poles. Because C_{μ} and C_{GD} in the CE/CS configuration introduce out-of-phase feed-forward paths to the output, they also introduce right-half-plane zeros. Capacitors C_{π} and C_{GS} in CC/CD followers, on the other hand, introduce in-phase feed-forward paths to the output (i.e., left-half-plane zeros) that offset the effects of their respective output poles (i.e., p_O). With respect to emitter/source impedance, C_{π} and C_{GS} also exert a polelike influence on transconductor current i_{gm} because they shunt v_{be} and v_{gs}, which means i_{gm} decreases and $1/g_m$ increases at higher frequencies. In other words, the influence these capacitors have on $1/g_m$ in CB/CG buffers produce an inductive effect on Z_{IN} in the same manner they do on Z_{OUT} in CC/CD followers.

In determining the gain of any circuit, when decomposed to its constituent single-transistor stages, it is helpful to view the circuit as a series of impedances through which collector/drain currents flow. Generally, transconductor collector/drain currents depend on base-emitter/gate-source voltages, and the manner in which incoming signals (i.e., voltages or currents) reach the base/gate terminals determines the nature of that voltage. As a result, the product of the resulting current and the total impedance at any point determines its voltage and the ratios of the resulting voltages and/or currents represent the respective gains of the circuit.

Figure 3.12 illustrates a graphical summary of all small-signal impedances present in a transistor circuit and their corresponding currents, including the v_{BE}- and v_{GS}-degenerating effects of impedances in series with the emitter/source and base/gate terminals on transconductances and input/output impedances. Although the gate-source resistance of a MOSFET is infinitely high, C_{GS} presents gate-source impedance Z_{GS} so the figure includes the effects Z_{GS} has on the rest of the circuit. Incidentally, including Z_{GS} in the model (except for bulk effects) equates the small-signal relationships that describe the small-signal behavior of bipolar and MOS transistors.

Just as series emitter and source impedances Z_E and Z_S degenerate g_m when driven from the base or gate, Z_B and Z_G also degenerate g_m when driven from the emitter or base because, as with emitter and source degeneration, only a voltage-divided fraction of the input voltage remains across the base-emitter or gate-source terminals. Z_G does not degenerate bulk-effect transconductance g_{mb}, however, because g_{mb}'s resulting current is independent of the gate voltage, which means g_{mb} sees the entire source voltage and not a fraction of it. In any case, although the exhaustive summary seems a bit overwhelming, it is a succinct reminder of the degenerating and loading effects of impedances in the circuit on various performance parameters. The figure becomes even more practical when used in parts, that

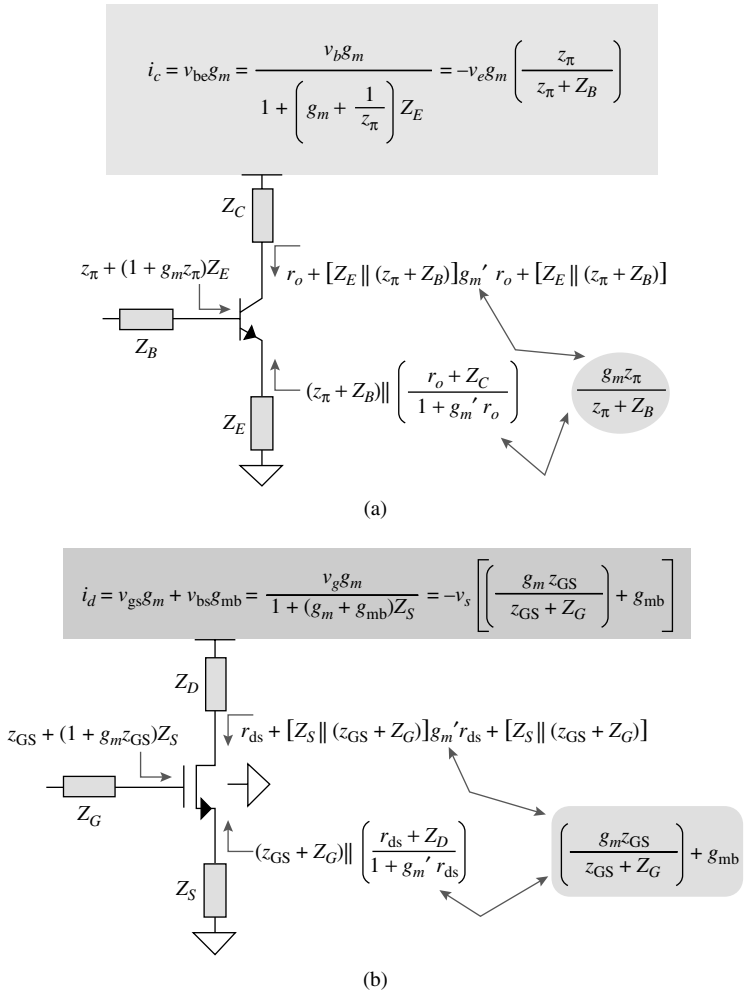


Figure 3.12 (*a*) NPN and (*b*) MOS transistors and their respective impedances and currents.

is, when (1) first analyzing low-frequency response and simplifying all impedances to resistances; (2) applying technology- and circuit-specific assumptions such as r_π is considerably greater than $1/g_m$, g_{mb} disappears when source and bulk terminals are short-circuited, and so on; and so forth.

A relatively straightforward approach of ascertaining gains across a circuit is to decompose the signals into their current and impedance equivalents and calculate the ohmic effects of the former on the latter as incoming signals traverse through the circuit. Consider transistor Q_N in the CE gain stage shown in Fig. 3.13 and the

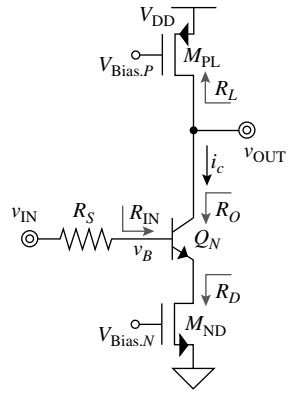


FIGURE 3.13
Illustrative example on the application of the impedance and transconductance summary offered in Fig. 3.12 to determine the constituent currents and voltages that make up the various gains of the circuit.

effects of source resistance R_S, degenerating transistor M_D, and load M_L on low-frequency small-signal voltage gain A_{V0} or v_{out}/v_{in}. First, voltage divider R_S-R_{IN} attenuates incoming small-signal input voltage v_{in} and impresses its result on Q_N as base voltage v_b. This voltage, in turn, modulates collector current i_c via Q_N's degenerated transconductance G_M—note R_D or M_D's $r_{ds.D}$ degenerates Q_N's $g_{m.N}$. Finally, Q_N pulls i_c from the equivalent output resistance present at v_{out} (i.e., collector voltage is negative), where the resistance is the parallel combination of R_L or M_L's $r_{sd.L}$ and Q_N's R_O. As a result, overall low-frequency, small-signal voltage gain A_{V0} or v_{out}/v_{in} is the product of the gains from v_{in} to v_b, v_b to i_c, and i_c to v_{out}:

$$\begin{aligned}
\frac{v_{out}}{v_{in}} &= \left(\frac{v_b}{v_{in}}\right)\left(\frac{i_c}{v_b}\right)\left(\frac{v_{out}}{i_c}\right) = \left(\frac{R_{IN}}{R_S + R_{IN}}\right)(-G_M)(R_L \,\|\, R_O) \\
&= \left(\frac{r_\pi + (1+\beta)R_D}{R_S + r_\pi + (1+\beta)R_D}\right)\left[\frac{-g_m}{1+\left(g_m + \frac{1}{r_\pi}\right)R_D}\right] \\
&\quad \times (R_L \,\|\, \{[(r_\pi + R_S) \,\|\, R_D] + r_o + g_m[(r_\pi + R_S) \,\|\, R_D]r_o\}) \\
&\approx \left(\frac{\beta R_D}{R_S + \beta R_D}\right)\left(\frac{-g_m}{1+g_m R_D}\right)(R_L \,\|\, \{g_m[(r_\pi + R_S) \,\|\, R_D]r_o\}) \\
&\approx (1)\left(\frac{-1}{R_D}\right)R_L = -\left(\frac{r_{sd.L}}{r_{ds.D}}\right)
\end{aligned} \tag{3.46}$$

assuming β is much larger than one, βR_D overwhelms r_π and R_S, and R_L is considerably smaller than R_O, which are all reasonable but not always necessarily so.

3.2 Differential Pairs

A differential pair is an *emitter/source-coupled pair* of transistors, as illustrated in Fig. 3.14*a*, with a *tail current* source tied to their common emitter/source point, the purpose of which is to force the sum of the currents flowing through the transistors to remain constant. Because there are two transistors and two currents, there are also two collector/drain output terminals. The average of the signals present at the bases/gates of the transistors constitutes the common portion of their respective inputs (i.e., *common-mode voltage* v_{CM}) and the difference the differential part (i.e., *differential voltage* v_{ID}). Differential voltage v_{ID} splits equally between the input terminals: the positive half (i.e., 0.5 v_{ID}) to one and the negative half (i.e., –0.5 v_{ID}) to the other.

Because the intrinsic inputs are voltages and outputs are currents, the differential pair, in its most basic form, is a transconductor because it converts a voltage into a current. Applying only a common-mode signal to the differential pair, however, irrespective of its value, results in equally split currents (at 0.5 I_{Bias}) because its base-emitter or gate-source voltages equal (that is, bases or gates are at v_{CM} and emitters or sources are tied together). A differential voltage, on the other hand, induces different base-emitter or gate-source voltages and results in asymmetric output currents. It is because differential signals incur variations in the output and common-mode signals do not that the differential pair derives its name, as it processes v_{ID} and rejects v_{CM}. This ability to only process differential signals does not only avail a niche application space for the circuit but also provides the foundation for rejecting common-mode noise in a system whose unmitigated effects can negate the performance and integration density achievements of modern-day, common-substrate technologies, being that a common substrate is a conductive medium for noise.

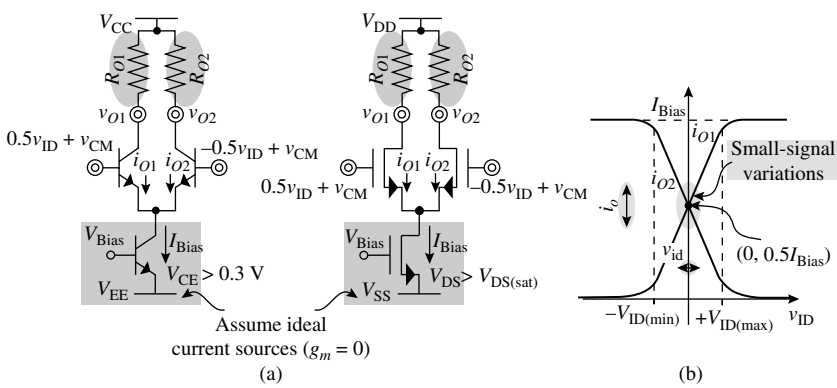


Figure 3.14 (*a*) Differential pair and (*b*) its large-signal response.

Large-Signal Operation

An important feature of the differential pair is its constant tail current (e.g., I_{Bias}) because it relates the effects of one output current onto the other. For instance, applying a positive differential voltage v_{ID} (Fig. 3.14*a*) increases the output current in one transistor by the same amount it decreases the current of the other so that, as output current i_{O1} increases, its counterpart i_{O2} decreases at an equal rate and by an equal amount, as shown in Fig. 3.14*b*. When $|v_{\text{ID}}|$ is large enough to fully steer the tail current (e.g., I_{Bias}) through one transistor, the differential pair saturates (e.g., I_{O1} to I_{Bias} and I_{O2} to zero, or vice versa) and can no longer incur variations in its output currents, giving rise to the symmetrical "S" curves shown in the large-signal response of Fig. 3.14*b*. This saturation point marks the linear-range limit of the circuit, that is to say, the maximum positive or negative differential voltage ($V_{\text{ID(max)}}$) that produces mostly linear variations in output currents and beyond which one transistor (and one current) carries all of the tail current (e.g., I_{Bias}):

$$V_{\text{ID(max).BJT}} \equiv \frac{\Delta i_{C(\max)}}{g_m} \approx \frac{I_{\text{Bias}}}{g_m} \approx \frac{I_{\text{Bias}} V_t}{0.5 I_{\text{Bias}}} = 2V_t \tag{3.47}$$

assuming $V_{\text{ID(max).BJT}}$ is sufficiently small to be approximated with linear, first-order small-signal models and V_t is the thermal voltage, which is approximately 25.6 mV at room temperature. Similarly, in the MOS case, $V_{\text{ID(max).MOS}}$ is approximately

$$V_{\text{ID(max).MOS}} \equiv \frac{\Delta i_{D(\max)}}{g_m} \approx \frac{I_{\text{Bias}}}{g_m} \approx \frac{I_{\text{Bias}}}{\sqrt{2(0.5I_{\text{Bias}})K'\left(\frac{W}{L}\right)}} = \sqrt{\frac{2(0.5I_{\text{Bias}})}{K'\left(\frac{W}{L}\right)}} = V_{\text{DS(sat)}} \tag{3.48}$$

where $V_{\text{ID(max).MOS}}$ is 100 mV if I_{Bias}, K', and (W/L) are 50 μA, 100 μA/V^2, and 50, respectively. Note differential voltages necessarily range in the sub-milli-volt region, unless the effective transconductances of the devices are degenerated, as shown in Fig. 3.15, where

$$V_{\text{ID(max).DEG}} \equiv \frac{\Delta i_{D(\max)}}{g_m} \approx \frac{I_{\text{Bias}}(1 + R g_m)}{g_m} \geq V_{\text{ID(max)}} \tag{3.49}$$

which when using similar numbers and a degenerating resistance of 50 kΩ, is approximately 2.55 V and 2.6 V in the BJT and MOS cases, respectively. (Refer to the low-frequency small-signal subsection of a degenerated CE/CS transistor for a discussion on transconductance g_m and its degenerated equivalent.)

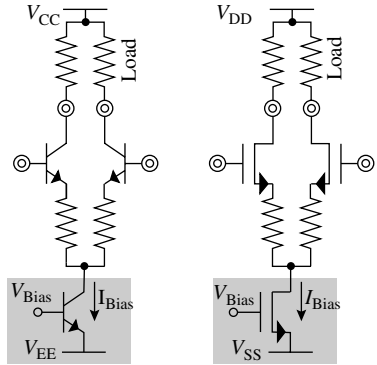


Figure 3.15 Emitter/source degenerated npn and n-type MOS differential pairs.

The differential pair may not process common-mode voltage v_{CM} but arbitrarily extending v_{CM}'s dc value (i.e., V_{CM}) ultimately forces one or several transistors in the circuit out of their respective high-gain regions (i.e., deep into saturation for the BJT and triode for the MOSFET), degrading the circuit's overall gain performance. For instance, in an n-type differential pair, reducing V_{CM} decreases the common emitter- or source-coupled node voltage (at a V_{BE} or V_{GS} below it), subjecting the tail-current bias circuit to even lower voltages, the extreme of which degrades I_{Bias} by pushing one or more of its defining transistors into triode. Similarly, higher V_{CM} voltages pull the common emitter-/source-coupled node to higher voltages, the extreme of which, when pressed against the load, pushes one or both transistors in the differential pair into their low-gain regions (i.e., triode). As a result, the minimum voltage across the tail current (i.e., $V_{Tail(min)}$) and the maximum voltage across the load (i.e., $V_{Load(max)}$) limit the operational *input common-mode range* (*ICMR*) of the circuit:

$$V_{EE(max)} + V_{Tail(min)} + V_{BE(max)} < V_{CM} < V_{CC(min)} - V_{Load(max)} - V_{CE(min)} + V_{BE(min)} \tag{3.50}$$

where $V_{Tail(min)}$ for a CE/CS transistor refers to the transistor's deep saturation limit $V_{CE(min)}$, $V_{CE(min)}$ is approximately 0.2–0.3 V, and $V_{SS(max)}$, $V_{GS(max)}$, $V_{DD(min)}$, $V_{DS(sat)}$, and $V_{GS(min)}$ replace $V_{EE(max)}$, $V_{BE(max)}$, $V_{CC(min)}$, $V_{CE(min)}$, and $V_{BE(min)}$, respectively, for the MOS case. Note worst-case extremes of V_{BE} and V_{GS} occur at the extreme range limits of tail current source I_{Bias} and temperature.

Although connecting the bulk to the source circumvents bulk effects, doing so exposes the source to common substrate noise, and connecting the bulk to one of the supplies does not. Tying the bulks to the supply, however, increases the effective threshold voltage (V_T) of the input pair. This increase in V_T decreases ICMR against the supply with the tail-current transistor because a higher margin must now exist between V_{CM} and the supply before the tail-current transistor

transitions into triode. Interestingly, the same increase in V_T also extends ICMR against the supply carrying the load because a smaller margin must now exist between V_{CM} and the load before one or both of the differential pair transistors transition into triode.

3.2.1 Differential Signals

Small-Signal Response at Low Frequency

In ascertaining the small-signal ac performance of the differential pair, it is helpful to decouple differential ac input signal v_{id} and common-mode ac input signal v_{cm}. In treating them separately, one is analyzed and the other assumed zero (and vice versa), and coupled back through superposition. In the case of differential signals only, v_{cm} is zero and, because equal but opposite fractions of the differential voltage are applied to the input terminals of the differential pair (and tail current source I_{Bias} remains constant), output currents i_{o1} and i_{o2} are also equal and opposite (i.e., $i_{o1} = -i_{o2}$ and $|i_{o1}| = |i_{o2}| = |i_o|$, as denoted in Fig. 3.16). The resulting common emitter/source ac voltage, which is the ohmic drop across the equivalent output resistance of the tail current (i.e., R_{Tail}), is zero because the net current flow into R_{Tail} is zero. Consequently, the common emitter/source node becomes a *virtual ac ground* for differential signals. Note, however, this is only

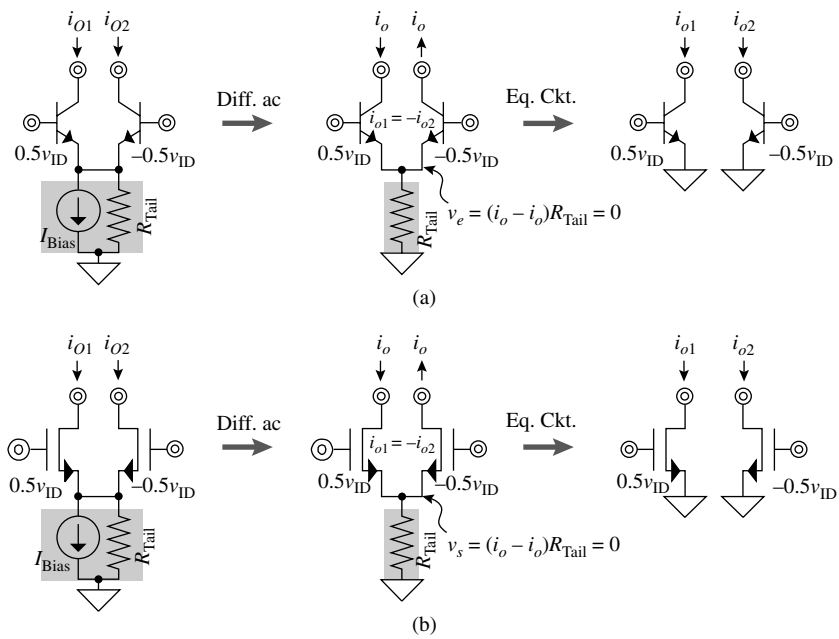


Figure 3.16 Differential large- to small-signal equivalent circuit transformations for (*a*) npn and (*b*) n-type MOS differential pairs.

true for differential signals wherein equal and opposite fractions of v_{id} are at the respective input terminals of the circuit.

The small-signal equivalent circuit of the differential pair for differential signals reduces to two CE/CS transistor amplifiers, each with half the input signal. Their respective small-signal parameters equal because their biasing conditions (i.e., collector/drain currents and base-emitter/gate-source voltages) and physical properties equal—both transistors are, by design, matched. Consequently, the differential gain across each transistor (i.e., A_{D1} and A_{D2}), because its base/gate's small-signal voltage only has half the input, is half the gain of a basic CE/CS gain stage:

$$A_{D1} \equiv \frac{v_{o1}}{v_{id}} = \frac{0.5v_{id}(-g_m)(R_{O1} \| r_o)}{v_{id}} = -0.5g_m(R_{O1} \| r_o) \tag{3.51}$$

and

$$A_{D2} \equiv \frac{v_{o2}}{v_{id}} = \frac{-0.5v_{id}(-g_m)(R_{O2} \| r_o)}{v_{id}} = 0.5g_m(R_{O2} \| r_o) \tag{3.52}$$

where r_{ds} replaces r_o in the MOS case. Note the gain can be inverting or noninverting depending on the output selected so the circuit is flexible in this respect. Further processing these outputs in a differential fashion and ensuring their respective loading resistances equal (i.e., $R_{O1} = R_{O2} \equiv R_{Load}$) increases their effective gain by two and produces the same gain as the basic CE/CS circuit:

$$A_{DD} \equiv \frac{v_{od}}{v_{id}} \equiv \frac{v_{o2} - v_{o1}}{v_{id}} = \frac{v_{o2}}{v_{id}} - \frac{v_{o1}}{v_{id}} = g_m(R_{Load} \| r_o) \tag{3.53}$$

where A_{DD} is the differential-to-differential gain and v_{od} the effective differential output voltage. In the emitter/source-degenerated case of Fig. 3.15, the degenerated transconductance replaces g_m in the aforementioned gain relationships.

Differential input resistance R_{ID} is the resistance across the input terminals of the differential pair, which is a measure of how much small-signal input current flows between the input terminals when applying a differential voltage. Since half the differential voltage is present across each CE/CS base-emitter terminal, half the current flows into each base terminal and twice the resistance results, when compared to the basic CE/CS stage:

$$R_{ID} \equiv \frac{v_{id}}{i_{id}} = \frac{v_{id}}{\left(\frac{0.5v_{id}}{r_\pi}\right)} = 2r_\pi \tag{3.54}$$

Because the MOS has no gate current, R_{ID} is infinitely large in the MOS case. Similarly, differential output resistance R_{OD} is a measure of the effective current flow across the output terminals of the differential pair in the presence of a differential output voltage v_{od}. Because only half the full differential output voltage v_{od} is present at each output terminal (i.e., $|v_{o1}| = |v_{o2}| = |0.5v_{od}|$), R_{OD} is effectively twice that of the basic CE/CS circuit:

$$R_{OD} \equiv \frac{v_{od}}{i_{od}} = \frac{v_{od}}{\left(\frac{0.5v_{od}}{R_{O1}}\right)} = 2(R_{Load} \,||\, r_o) \tag{3.55}$$

where r_{ds} replaces r_o in the MOS case. If only one output terminal is used (i.e., a single-ended output), its output resistance is that of the basic CE/CS transistor ($R_{Load} \,||\, r_o$ or $R_{Load} \,||\, r_{ds}$).

Small-Signal Response at High Frequency

Because the differential pair in differential mode decomposes into CE/CS stages, its frequency response mimics that of the CE/CS amplifier. As such, base-emitter C_π and Miller-multiplied base-collector C_μ capacitors (or gate-source C_{GS} and Miller-multiplied gate-drain C_{GD} capacitors) produce the effects of a pole at each input. Capacitor C_μ or C_{GD} also offer out-of-phase, feed-forward paths and therefore yield the effect of a right-half-plane zero through each input terminal. At frequencies beyond which C_μ's or C_{GD}'s effectively short their equivalent parallel resistances, the output resistance of the circuit reduces to $1/g_m$ because their respective outputs are now impressed on their inputs (as diode-connected transistors). Loading capacitor C_{Load} then shunts $1/g_m$, producing the effects of an output pole at each output terminal. In the case where C_μ's or C_{GD}'s do not short until higher frequencies (when source resistance R_S is low), output pole p_O asserts its influence before p_{IN}, when the parasitic capacitance present at the output shorts the output resistance:

$$Z_{C.PAR} \equiv \left.\frac{1}{s(C_\mu + C_{Load})}\right|_{p_O \approx \frac{1}{2\pi(r_o||R_{Load})(C_\mu + C_{Load})}} \equiv r_o \,||\, R_{Load} \tag{3.56}$$

where $Z_{C.PAR}$ is the impedance across the capacitors and C_μ or C_{GD} and C_{LOAD} constitute the output capacitance.

3.2.2 Common-Mode Signals

Small-Signal Response at Low Frequency

In developing an equivalent small-signal common-mode circuit for the differential pair, differential voltage v_{ID} is zero and common-mode

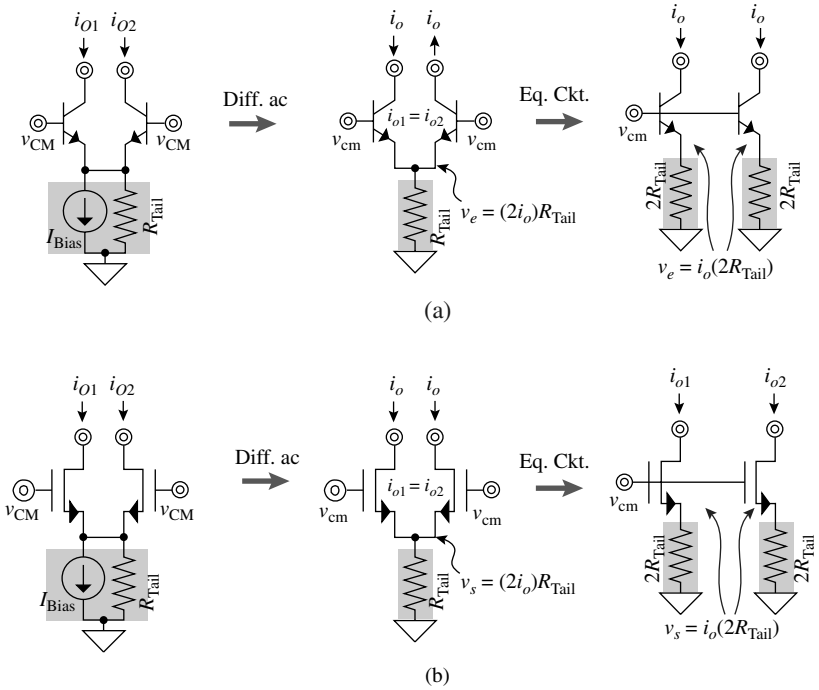


FIGURE 3.17 Common-mode large to small-signal equivalent circuit transformations of (*a*) npn and (*b*) n-type MOS differential pairs.

signal v_{CM} is alone considered. Because both base or gate voltages equal and their emitter or source terminals share a common node (and their voltages equal), their respective output currents also equal (i.e., $i_{o1} = i_{o2} \equiv i_o$). As a result, the small-signal voltage present at their common emitter/source node is the ohmic drop across tail current source resistance R_{Tail} induced by said small-signal currents [i.e., $(2i_o)R_{Tail}$]. Note, however, the same emitter/source voltage (i.e., same level of degeneration) results from having one output current flow through twice the tail current-source resistance [i.e., $i_o(2R_{Tail})$], as shown in the transformations of Fig. 3.17. This transformation is useful because it converts the differential pair into two emitter/source-degenerated CE/CS transistors in parallel, the analysis of which preceded this section.

The small-signal gain across each transistor in the pair is therefore the emitter/source-degenerated gain of a CE/CS stage:

$$A_{C1} = A_{C2} \equiv \frac{v_{o1}}{v_{cm}} \approx \frac{v_{cm}\left(\dfrac{-g_m}{1+2R_{Tail}g_m}\right)(R_{O1} \| r_o)}{v_{cm}} = -\left(\frac{g_m}{1+2R_{Tail}g_m}\right)(R_{O1} \| r_o) \tag{3.57}$$

where A_{C1} and A_{C2} are the common-mode to single-ended gains and r_{ds} replaces r_o in the MOS case. Unlike the differential counterpart, both gains are in-phase so further processing the outputs in a differential fashion and ensuring their respective loading resistances equal (i.e., $R_{O1} = R_{O2} \equiv R_{Load}$) cancels their common-mode gains to zero:

$$A_{CD} \equiv \frac{v_{od}}{v_{cm}} \equiv \frac{v_{o2} - v_{o1}}{v_{id}} = \frac{v_{o2}}{v_{id}} - \frac{v_{o1}}{v_{id}} = 0 \tag{3.58}$$

where A_{CD} is the common-mode to differential gain of the circuit. This is ideal because the differential output rejects all common-mode noise present. A measure of how much the differential pair favors differential over common-mode signals is *common-mode rejection ratio* (*CMRR*), which is ideally infinitely large:

$$\text{CMRR} \equiv \frac{A_{DD}}{A_{CD}} = \frac{A_{DD}}{0} \to \infty \tag{3.59}$$

In practice, however, there are mismatches between loading resistors R_{O1} and R_{O2} and the input transistors of the differential pair, resulting in mismatched common-mode gains A_{C1} and A_{C2} and therefore finite CMRRs. Because mismatches are relatively small in modern-day technologies, however, CMRR is still relatively high (e.g., 80 dB). Even if only a single output is processed and differential cancellation at the output is no longer available, CMRR (or in this case A_{DD}/A_{C1}) can still be moderately high because common-mode gains A_{C1} and A_{C2} are emitter/source-degenerated with a relatively large tail-current source resistance (i.e., $2R_{Tail}$), which is at worst on the order of r_o or r_{ds} (e.g., 0.5–1 M Ω).

The common-mode input resistance of the differential pair (i.e., R_{IC}) is the parallel combination of the input resistances of two emitter/source-degenerated CE/CS circuits (i.e., $R_{IN.DEG}$), which is half the input resistance of one and often approximates to βR_{Tail} in the BJT case:

$$\begin{aligned} R_{IC} \equiv R_{IN.DEG} \parallel R_{IN.DEG} &= 0.5R_{IN.DEG} \\ &= 0.5[r_\pi + (1+\beta)(2R_{Tail})] \approx \beta R_{Tail} \end{aligned} \tag{3.60}$$

and approaches infinity in the MOS case. The output resistance of each terminal (i.e., R_{OC1} and R_{OC2}) is that of the emitter/source-degenerated CE/CS circuit, except the degenerating resistance is now $2R_{Tail}$, which itself is relatively large to begin with:

$$R_{OC1} = R_{OC2} \approx g_m r_o (2R_{Tail} \parallel r_\pi) \approx \beta r_o \tag{3.61}$$

where r_π disappears and R_{OC1} and R_{OC2} no longer reduce to βr_o but remain high in the MOS case.

Small-Signal Response at High Frequency

The high-frequency response of common-mode signals mimics that of the degenerated CE/CS circuit, which is mostly a derivative of its basic CE/CS case. As such, C_μ or C_{GD} introduce a Miller-multiplied pole at the input and a right-half-plane zero across the transistors, loading capacitor C_{Load} constitutes another pole at the output, and the parasitic capacitance across degenerating resistor $2R_{Tail}$ introduces a zero-pole doublet. Because the inverting gain is lower than the nondegenerated case and the common-mode source resistance is often low, the Miller-defined input pole often exceeds output pole p_O so the latter asserts its effects at lower frequencies when the load capacitance and C_μ or C_{GD} short output resistance $R_{Load} \,||\, R_{OC}$:

$$Z_{C.PAR} \equiv \left.\frac{1}{s(C_\mu + C_{Load})}\right|_{p_O \approx \frac{1}{2\pi(R_{Load}||R_{OC})(C_\mu + C_{Load})}} \equiv R_{Load} \,||\, R_{OC} \qquad (3.62)$$

Note a zero introduced by common-mode gain A_C (i.e., z_X) constitutes a pole in CMRR, which means CMRR degrades past any zero in A_C:

$$\text{CMRR} \equiv \frac{A_D}{A_C} = \frac{A_{D.LF}(\ldots)}{A_{C.LF}\left(1 + \frac{s}{z_G}\right)(\ldots)} \qquad (3.63)$$

where A_D refers to the differential gain and subscript "LF" to low frequency. Unfortunately, because tail resistance R_{Tail} is considerably large and the parasitic capacitance present (i.e., C_{Tail}) is not negligibly small, especially in the presence of source-bulk or bulk-substrate capacitances, degenerated transconductance G_M's zero (i.e., z_G) is at relatively low frequencies:

$$Z_{C.Tail} \equiv \left.\frac{1}{sC_{Tail}}\right|_{z_G = \frac{1}{2\pi R_{Tail} C_{Tail}}} \equiv R_{Tail} \qquad (3.64)$$

Incidentally, C_{Tail} splits in the common-mode equivalent half circuit, as shown in Fig. 3.18. Also note p_O is often dominant and set by R_{Load} (i.e., R_{Load} is considerably smaller than R_{OUT}) in both differential and common-mode cases so their effects on CMRR tend to cancel and CMRR still degenerates past z_G.

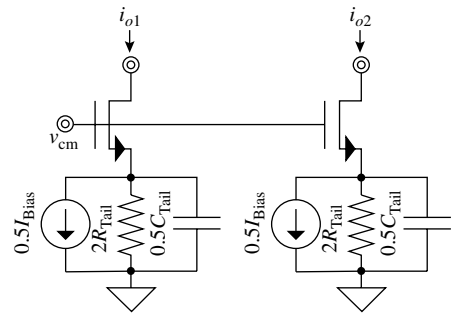


FIGURE 3.18
Common-mode equivalent half circuit.

3.3 Current Mirrors

3.3.1 Basic Mirror

Operation

The objective of a current mirror is to duplicate its input current at its output, irrespective of the output voltage present. To do this without errors, two transistors must match and all their respective port-to-port voltages must equal. However, since collector-emitter and drain-source voltages v_{CE} and v_{DS} have minimal impact on collector and drain currents i_C and i_D in the transistors' high-gain mode (i.e., forward active or slightly saturated for BJTs and saturation for FETs), v_{CE} and v_{DS} need not equal for their currents to match reasonably well, as long as v_{BE}'s and v_{GS}'s equal. For example, short-circuiting the base-emitter and gate-source control terminals of matched transistors, as shown in Fig. 3.1*a* and *b*, and ensuring their respective collector-emitter and drain-source voltages are sufficiently high, produce an output current that is approximately equal to its input:

$$i_C = I_S \exp\left(\frac{v_{BE}}{V_t}\right)\left(1+\frac{v_{CE}}{V_A}\right) \approx I_S \exp\left(\frac{v_{BE}}{V_t}\right) \approx i_{IN} \approx i_{OUT} \tag{3.65}$$

and

$$i_D = \left(\frac{W}{L}\right)K'(v_{GS}-V_T)^2(1+\lambda v_{DS}) \approx \left(\frac{W}{L}\right)K'(v_{GS}-V_T)^2 \approx i_{IN} \approx i_{OUT} \tag{3.66}$$

Connecting the input transistor's collector (or drain) to the base (or gate) ensures the input current charges the parasitic base-emitter (or gate-source) capacitance until the collector (or drain) current is equal to input current i_{IN}, at which point the capacitor stops charging

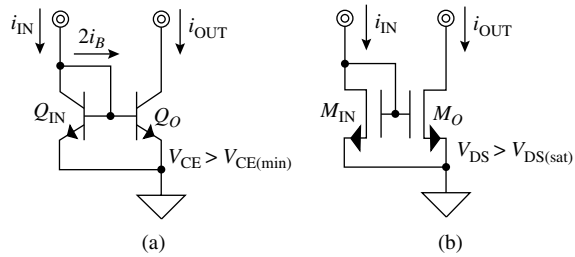


Figure 3.19 Basic (*a*) npn and (*b*) n-type MOS current mirrors.

and v_{BE} (or v_{GS}) is set. Because connecting the collector and base terminals amounts to using the BJT as a base-emitter pn-junction diode, Q_{IN} and M_{IN} in Fig. 3.19*a* and *b* are said to be diode connected.

This mirror configuration is also useful for amplifying or attenuating a current by a constant factor because changing the number of transistors in the input and/or output of the circuit modifies the circuit's input-output gain ratio. For instance, using three matched parallel transistors in the input equally splits input current i_{IN} into three and using two matched parallel transistors at the output ensures the fractional current flowing through each input device (i.e., $i_C = i_{IN}/3$) flows through each output device, yielding a sum total of $2i_C$ or $2i_{IN}/3$, as depicted in Fig. 3.20*a*. This mirror configuration mimics the effects of having an input transistor with 3 times the minimum emitter area and an output device with 2 times the minimum emitter area. The same idea applies to MOSFETs, except gain is set by width-length aspect ratios (i.e., W/L), as illustrated in the 2× mirror shown in Fig. 3.20*b*.

Performance

Important metrics of a current mirror include its accuracy, input and output voltage limits, input and output resistances, gain, and bandwidth. To start, with respect to accuracy, base currents in BJTs subtract a fraction of the input current flowing into the mirror. In the case of a one-to-one npn mirror, for example, bases pull and subtract two base

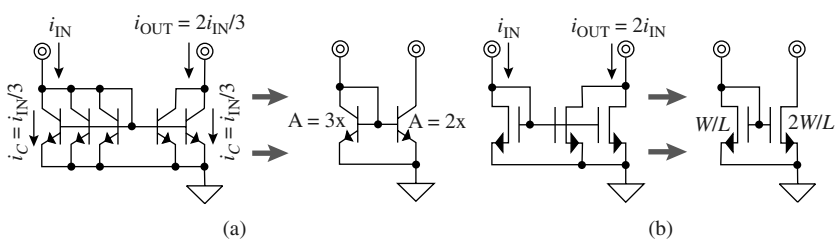


Figure 3.20 Nonunity (*a*) BJT and (*b*) MOSFET current mirrors.

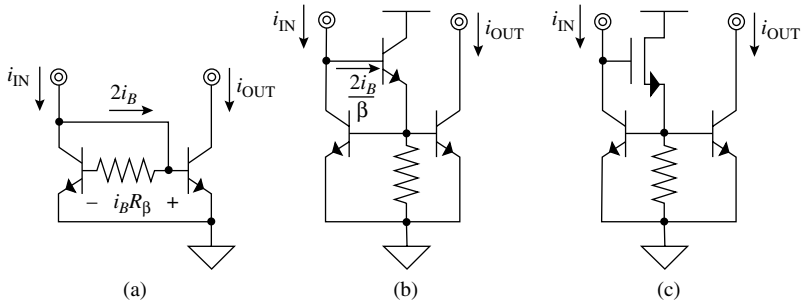


FIGURE 3.21 Basic npn current mirrors with (*a*) β-compensating resistor, (*b*) npn β-helper, and (*c*) MOS β-helper.

currents (i.e., $2i_B$) from input current i_{IN} and consequently cause the output (which is a collector) to sink a lower current (i.e., $i_{OUT} < i_{IN}$),

$$i_{IN} \approx i_C + 2i_B \approx i_C\left(1+\frac{2}{\beta}\right) = i_{OUT}\left(1+\frac{2}{\beta}\right) \tag{3.67}$$

where i_B is the base current into Q_{IN} and Q_O in Fig. 3.19*a*. Slightly increasing the base-emitter voltage of output transistor Q_O with an ohmic voltage drop increases its collector current, partially compensating for the mirror's β error, as illustrated in Fig. 3.21*a*. Since the collector's current variation Δi_{OUT} is the result of a small change in its base-emitter voltage Δv_{BE}, equating resistor voltage v_R to the linear first-order transconductance (i.e., g_m) translation of the error current (i.e., Ni_B, where N is the effective number of bases attached carrying equivalent input currents) compensates the mirror error to first order:

$$\Delta i_{OUT} = Ni_B \equiv \Delta v_{BE} g_m = v_R g_m = (i_B R_\beta) g_m \tag{3.68}$$

or

$$R_\beta \equiv \frac{N}{g_m} \approx \frac{NV_t}{I_{IN}} \tag{3.69}$$

where I_{IN} is the dc value of input current i_{IN}.

Alternatively, while placing a series β-helper BJT level shifter in the collector-base connection path, as shown in Fig. 3.2*b*, decreases the error current by a factor of β, a series MOSFET (Fig. 3.21*c*) altogether eliminates it, given the latter carries no gate current. The purpose of the load resistor attached to the β-helper transistor, by the way, is to establish a bias current that pushes the pole associated with

that node to high frequencies. In any case, the simplicity of the β-compensating resistor often offsets its relatively poorer accuracy, given its resistance cannot easily track or adjust to a changing input current i_{IN}.

As already mentioned, all port-to-port terminals in the mirror circuit must equal (i.e., all base- and collector-emitter and gate-, drain-, and bulk-source voltages equal) to produce a quasi-perfect replica of the input current at the output. The collector-emitter and drain-source voltages in the basic current mirror, however, do not match, and although their effects are minimal, they still induce a gain error that is the result of base-width and channel-length modulation effects:

$$A_I = \frac{i_{C.OUT}}{i_{C.IN}} = \frac{I_S \exp\left(\frac{v_{BE}}{V_t}\right)\left(1+\frac{v_{IN}}{V_A}\right)}{I_S \exp\left(\frac{v_{BE}}{V_t}\right)\left(1+\frac{v_{OUT}}{V_A}\right)} = \frac{1+\frac{v_{BE}}{V_A}}{1+\frac{v_{OUT}}{V_A}} \tag{3.70}$$

and

$$A_I = \frac{i_{D.OUT}}{i_{D.IN}} = \frac{\left(\frac{W}{L}\right)K'(v_{GS}-V_T)^2(1+\lambda v_{IN})}{\left(\frac{W}{L}\right)K'(v_{GS}-V_T)^2(1+\lambda v_{OUT})} = \frac{1+\lambda v_{GS}}{1+\lambda v_{OUT}} \tag{3.71}$$

where A_I refers to the mirror gain of each equally sized transistor segment in the current mirror, which would otherwise be 1. For instance, the v_{CE}-induced systematic gain error of a basic mirror with its output at 3 V and an Early voltage V_A of 50 V, assuming a base-emitter voltage of 0.7 V, is approximately 4.3%, excluding random mismatch errors between supposedly matched transistors.

The input and output voltage limits for which the mirror maintains its prescribed gain is also important. Ideally, the input and output voltages should reach both positive and negative supplies without significant impact on the output current. However, the input voltage in a basic current mirror is one base-emitter voltage V_{BE} or gate-source voltage V_{GS} from ground, which can be on the order of 0.55–1.5 V, depending on process, size, temperature, and current density. β-helper transistors unfortunately increase the overhead by another V_{BE} or V_{GS}, unlike β-compensating resistors, which are relatively benign to the circuit in this respect, albeit with poorer accurate. The voltage overhead associated with the output is the lowest voltage the output transistor can tolerate across its collector-emitter or drain-source terminals without significantly altering its output current, which corresponds to deep saturation limit $V_{CE(min)}$ or saturation voltage $V_{DS(sat)}$ (e.g., 0.2–0.5 V).

Small-Signal Response at Low Frequency

When considering the small-signal gain of the circuit, accounting for the effects of the source and load is important. Before including these effects, however, it is best to simplify the circuit where possible. To this end, Fig. 3.22*a* illustrates the complete small-signal equivalent circuit of the basic npn current mirror shown in Fig. 3.19*a* and Fig. 3.22*b* shows its simplified but loaded counterpart. In diode-connecting input transistor Q_{IN}, for instance, the equivalent resistance into Q_{IN}'s transconductor $g_{m.IN}$ becomes $1/g_{m.M}$ because the current flowing through the diode-connected device is linearly proportional to its own terminal voltages. As such, the input resistance of the circuit (i.e., R_{IN}) reduces to $1/g_{m.IN}$, the lowest resistance in the parallel combination that R_{IN} comprises, which also includes base-emitter resistors $r_{\pi.IN}$ and $r_{\pi.O}$ and Q_{IN}'s output resistance $r_{o.IN}$:

$$R_{IN} = r_{\pi.IN} \parallel r_{\pi.O} \parallel r_{o.IN} \parallel \frac{1}{g_{m.IN}} \approx \frac{1}{g_{m.IN}} \tag{3.72}$$

Base-collector capacitor C_μ also disappears in the simplified circuit because its two terminals are short-circuited, which means C_μ's displacement (i.e., ac) current is zero (i.e., C_μ produces no effects in the

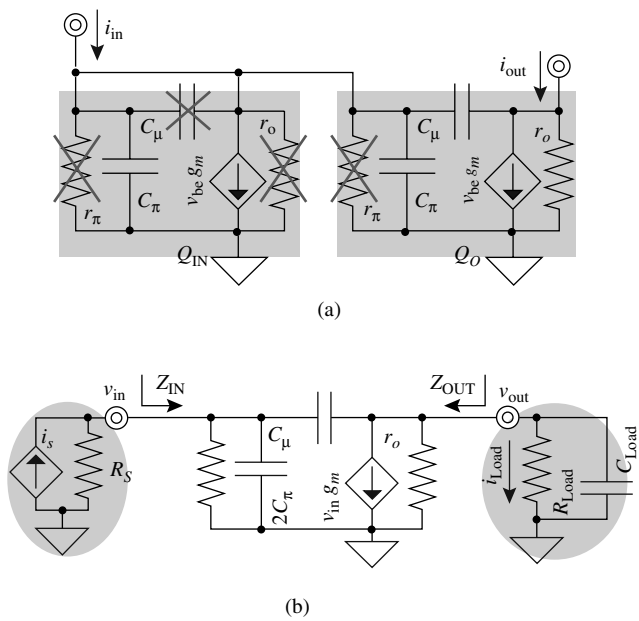


Figure 3.22 (*a*) Complete and (*b*) simplified (but loaded) small-signal model of a basic npn current mirror.

circuit). The output of the circuit remains intact and output resistance R_{OUT} reduces to Q_O's r_o, the only resistance present.

Perhaps the more meaningful definition of the small-signal gain of the mirror is the translation of source current i_s to load current i_{Load} in the presence of source resistance R_S and load capacitance C_{Load} because both R_S and C_{Load} represent realistic operating conditions. As discussed earlier in this chapter, decomposing the signal path into its equivalent current-voltage and voltage-current translations and using the results obtained in the single-transistor section helps ascertain the gain across the circuit. In this case, small-signal low-frequency current gain $A_{I.LF}$ is the product of small-signal translations i_s to input voltage v_{in}, v_{in} to transconductor current i_{gm}, i_{gm} to output voltage v_{out}, and v_{out} to i_{Load} or

$$A_{I.LF} \equiv \frac{i_{Load}}{i_s} = \left(\frac{v_{in}}{i_s}\right)\left(\frac{i_{gm}}{v_{in}}\right)\left(\frac{v_{out}}{i_{gm}}\right)\left(\frac{i_{Load}}{v_{out}}\right)$$

$$= (R_S \,||\, R_{IN})(-g_{m.O})(R_{OUT} \,||\, R_{Load})\left(\frac{1}{R_{Load}}\right)$$

$$= \left(R_S \,||\, \frac{1}{g_{m.IN}}\right)(-g_{m.O})(r_o \,||\, R_{Load})\left(\frac{1}{R_{Load}}\right) \approx -1 \qquad (3.73)$$

assuming a one-to-one area ratio between the matching transistors (i.e., $g_{m.IN}$ equals $g_{m.O}$). Note that a gain of 1 only results when R_{IN} is considerably smaller than R_S and R_{OUT} larger than R_{Load}; in other words, ideal input and output resistances R_{IN} and R_{OUT}, respectively, approach zero and infinitely high values.

Small-Signal Response at High Frequency

The operation described thus far assumes steady-state conditions and therefore relates to low frequency only. To ascertain the high-frequency response of the circuit, it is usually easier to start at low frequencies and find the poles that exist as frequency increases, short-circuiting the capacitor that produced poles at lower frequencies. Before analyzing the circuit, though, as noted earlier, Q_{IN}'s $C_{\mu.IN}$ is short-circuited (Fig. 3.22*a*) so no ac voltage exists across its terminals and consequently no ac current results through the device, which means, for all practical purposes, $C_{\mu.IN}$ does not exist.

Noting R_{IN} is generally low and R_{OUT} is high reveals output pole p_O precedes input pole p_{IN} as frequency increases. As a result, with respect to dominant pole p_O, loading capacitor C_{Load} and C_μ steer current away from the output when their collective impedance is equal to or smaller than their equivalent parallel resistance (i.e., $r_o \,||\, R_{Load}$). Because the base terminal side of C_μ is at a low-resistance point (i.e., $1/g_m$ from ground), its associated series impedance has negligible

effects on the total capacitor impedance so p_O occurs when both C_{Load} and C_μ shunt $r_o \,||\, R_{\text{Load}}$:

$$\frac{1}{C_{\text{Load}}s} \,||\, \left[\frac{1}{C_\mu s} + \left(\frac{1}{g_m} \,||\, R_S \,||\, \frac{1}{2C_\pi s}\right)\right] \approx \frac{1}{(C_{\text{Load}} + C_\mu)s}\bigg|_{p_O \approx \frac{1}{2\pi R_{\text{Load}}(C_{\text{Load}}+C_\mu)}} \equiv r_o \,||\, R_{\text{Load}} \approx R_{\text{Load}} \tag{3.74}$$

where C_{GD} and r_{ds} replace C_μ and r_o in the MOS case and s in Laplace transforms represents frequency.

As frequency increases past p_O, assuming C_{Load} is dominant, C_{Load} becomes a short circuit and Q_{IN}'s $C_{\pi.\text{IN}}$ and $C_{\mu.\text{IN}}$ and Q_O's $C_{\pi.O}$ begin to shunt and steer current away from Q_{IN}'s base, decreasing input voltage v_{in} and the output current it produces in transistor Q_O (i.e., i_{out}). The Miller effect of $C_{\mu.\text{IN}}$ is low because p_O, at this point, decreased much of Q_O's CE gain, which also means v_{out} is a low-impedance node (to ac ground). As a result, the effects of input (mirror) pole p_{IN} occur when equivalent input capacitance C_{IN} (or $C_{\pi.\text{IN}}$, $C_{\mu.O}$, and $C_{\pi.O}$) shunt the parallel combination of R_{IN} and R_S:

$$\frac{1}{sC_{\text{IN}}} \approx \frac{1}{s(C_{\pi.\text{IN}} + C_{\pi.O} + C_{\mu.O})}\bigg|_{p_{\text{IN}} \approx \frac{g_{m.\text{IN}}}{2\pi(C_{\pi.\text{IN}}+C_{\pi.O}+C_{\mu.O})}} \equiv R_{\text{IN}} \,||\, R_S \approx \frac{1}{g_{m.\text{IN}}} \tag{3.75}$$

where C_{GS} and C_{GD} replace C_π and C_μ and r_π disappears in the MOS case.

Capacitor C_μ or C_{GD}, as in the CE/CS transistor case, also constitute an out-of-phase feed-forward path whose effects prevail at and above frequencies where the capacitor current is equal to or greater than the mirroring transconductor current. Input pole p_{IN} precedes this right-half-plane zero (i.e., z_{RHP}), which resides at $g_m/2\pi C_\mu$ or $g_m/2\pi C_{\text{GD}}$, because the parasitic capacitance associated with the former exceeds the latter. It is important to keep z_{RHP} well above frequencies of interest because its right-half-plane effect not only extends bandwidth to problematic frequency regions, where several parasitic poles reside, but also decreases phase in the process.

3.3.2 Cascoded Mirror

Operation

The objective of the cascoded mirror with respect to its basic counterpart is twofold: (1) improve mirroring accuracy and (2) decrease sensitivity to variations in output voltage (i.e., increase output resistance).

The idea is to eliminate base-width and channel-length modulation effects in the mirroring devices by forcibly equating their respective collector-emitter or drain-source voltages with cascode transistors. Because cascode devices (e.g., Q_{CIN} and Q_{CO} or M_{CIN} and M_{CO} in Fig. 3.23*a* and *b*) carry the same current density, their base-emitter or gate-source voltages and therefore their emitter or source voltages equal (e.g., Q_{IN}'s and Q_O's collector-emitter voltages v_{CE}'s equal), eliminating all modulation effects. Since v_{CE}'s or v_{DS}'s are equal to v_{BE} or v_{GS} with respect to ground, variations in the output voltage have minimal impact on the port-to-port voltages of the mirroring devices (e.g., Q_{IN} and Q_O or M_{IN} and M_O), producing little effects in output current i_{OUT}, that is to say, increasing the output resistance of the mirror.

Performance

As in the basic mirror, base currents degrade the accuracy performance of cascoded mirrors and β-compensation resistors and β-helper transistors help reduce those effects. Unlike the basic mirror, however, base-width and channel-length modulation effects on the mirroring devices are less prevalent, although random, process-induced mismatch errors remain. Unfortunately, eliminating the modulation errors increases the input and output voltage requirements of the circuit. For instance, in the case of the cascoded mirrors in Fig. 3.23*a* and *b*, the dc input voltage is the voltage across two diode-connected devices (i.e., $2V_{BE}$ or $2V_{GS}$) and its output must exceed the diode-connected voltage impressed across the output mirroring device by one collector-emitter or drain-source voltage (i.e., $V_{BE} + V_{CE(min)}$ or $V_{GS} + V_{DS(sat)}$).

Relaxing the input and output headroom requirements of the cascode circuit amounts to eliminating the dc voltage effects of the cascode transistors. For instance, disconnecting the bases or gates of the cascodes from the circuit and appropriately biasing them one base-emitter or gate-source voltage above one saturation voltage, as in Fig. 3.24*a* and *b*, continues to equate the v_{CE}'s or v_{DS}'s across the

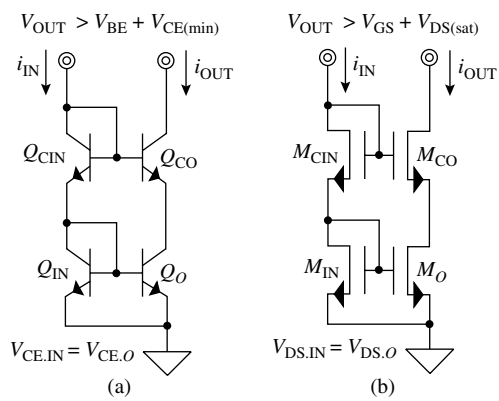


FIGURE 3.23
(*a*) NPN- and (*b*) NMOS-cascode current mirrors.

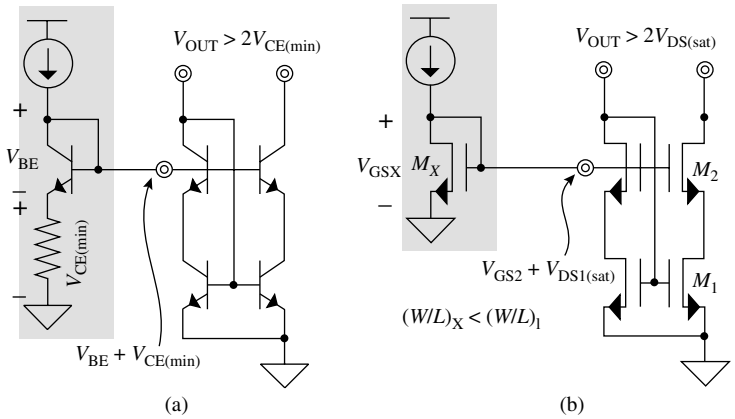


FIGURE 3.24 (*a*) NPN and (*b*) NMOS low-voltage cascode current mirrors.

mirroring devices (although now at $V_{CE(min)}$ or $V_{DS(sat)}$) while avoiding their V_{BE} or V_{GS} impact on headroom. The input voltage is therefore back to what it was in the basic current mirror, at one V_{BE} or V_{GS} above ground, but its output must now exceed two minimum saturation voltages, which is higher than in the basic mirror but lower than in the simple cascode version shown earlier. Note that generating the bias voltage requires additional circuit, silicon real estate, and power. In the sample BJT embodiment shown in Fig. 3.24*a*, for example, the resistor shifts up a base-emitter voltage by only the minimum collector-emitter voltage a BJT can sustain before entering the deep saturation region. Similarly, the aspect ratio (i.e., W/L) of biasing transistor M_X (and/or its biasing current) in the MOS example shown in Fig. 3.24*b* is sufficiently low to ensure its saturation voltage is large enough to accommodate the saturation voltage in the gate-source voltage of the cascode device (e.g., $V_{DS2(sat)}$) and ensure the voltage across the mirroring transistor is above its saturation point (e.g., $V_{DS1(sat)}$):

$$V_{GSX} = V_T + V_{DSX(sat)} \equiv V_{GS2} + V_{DS1(sat)} = V_T + V_{DS2(sat)} + V_{DS1(sat)} \qquad (3.76)$$

or

$$V_{DSX(sat)} \equiv V_{DS1(sat)} + V_{DS2(sat)} \qquad (3.77)$$

Small-Signal Response at Low Frequency

Figure 3.25 presents the complete and simplified but loaded small-signal equivalent circuits of the cascoded mirror circuit, which amount to two diode-connected devices on the input and a CE cascode

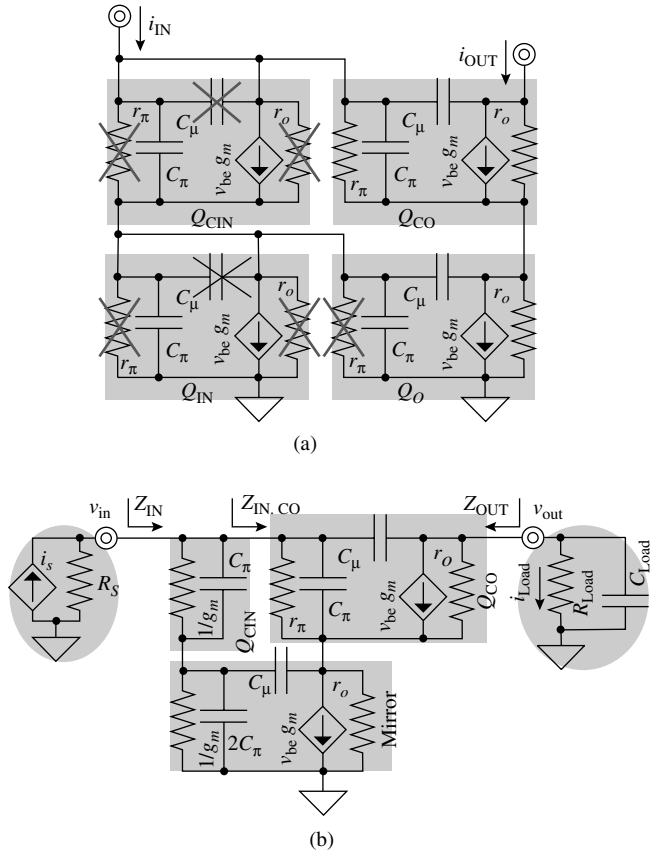


Figure 3.25 (*a*) Complete and (*b*) simplified (but loaded) small-signal equivalent circuits of the cascode current mirror.

transistor degenerated with a CE transistor on the output. Transistors Q_{IN} and Q_O constitute a basic mirror and they therefore simplify to the small-signal equivalent circuit shown in Fig. 3.22*b* and again in Fig. 3.25*b*. Similar to the mirror, the current flowing through Q_{CIN}'s base-emitter resistance $r_{\pi.CIN}$ and output resistance $r_{o.CIN}$ are negligibly smaller than their transconductor counterpart, whose equivalent resistance is $1/g_m$, so their effects, for all practical purposes, disappear in the simplified small-signal circuit. Q_{CIN}'s base-collector capacitor C_μ also disappears because there is no voltage across it to induce a current. As always, r_π's disappear and v_{gs}, C_{GS}, C_{GD}, and r_{ds} replace v_{be}, C_π, C_μ, and r_o, respectively, in the MOS case.

The input resistance (i.e., R_{IN}) of the circuit is the parallel combination of the resistance into the two diode-connected devices and the resistance into the base of output cascode transistor Q_{CO}. The resistance Q_{CO} presents, however, is considerably higher than the two $1/g_m$

resistances associated with the diode-connected devices because of the degenerating effects mirroring output resistance $r_{o.M}$ has on Q_{CO}. As a result, Q_{CO} introduces a β-translation of $r_{o.M}$ to the input, reducing R_{IN} to roughly $2/g_m$:

$$R_{IN} = \left(\frac{1}{g_{m.CIN}} + \frac{1}{g_{m.M}}\right) \| R_{IN.CO} = \left(\frac{1}{g_{m.CIN}} + \frac{1}{g_{m.M}}\right) \| [r_{\pi.CO} + (1+\beta)r_{o.M}] \approx \frac{2}{g_m} \tag{3.78}$$

where r_π disappears and β approaches infinity in the MOS case.

The degenerating effects of $r_{o.M}$ on Q_{CO} also manifest themselves in the output resistance of the circuit (i.e., R_{OUT}), except $r_{o.M}$ is not the only degenerating factor. In fact, Q_{CO}'s base-emitter resistance $r_{\pi.CO}$, the diode-connected devices ($2/g_m$), and the mirror's output transconductor $g_{m.M}$ present another resistance at Q_{CO}'s emitter. Because the bottom devices constitute a current mirror, output mirror transistor Q_O sinks a mirrored version of the base current flowing through output cascode device Q_{CO} (i.e., $i_{gm.O} \approx i_{\pi.CO}$), which means the total emitter-degenerating current is twice Q_{CO}'s base current and $r_{o.M}$'s smaller current contribution, which means the equivalent emitter degenerating resistance is approximately $r_\pi/2$:

$$R_{DEG.CO} \equiv \frac{v_e}{i_e} = \frac{v_e}{i_{\pi.CO} + i_{gm.O} + i_{o.M}} = \frac{v_e}{2i_{\pi.CO} + i_{o.M}} = \frac{v_e}{\left(\frac{2v_e}{r_\pi + \frac{2}{g_m}}\right) + \frac{v_e}{r_{o.M}}} \approx \frac{r_\pi + \frac{2}{g_m}}{2} \approx \frac{r_\pi}{2} \tag{3.79}$$

where $R_{DEG.CO}$ is Q_{CO}'s effective degenerating resistance, $i_{\pi.CO}$ is Q_{CO}'s base current, $i_{gm.O}$ is Q_O's transconductor current, and $i_{o.O}$ is the current flowing through $r_{o.M}$ (which is considerably smaller than $i_{\pi.CO}$). Note that in the MOS case r_π disappears and $i_{\pi.CO}$ is zero so degenerating resistance $R_{DEG.CO}$ reduces to $r_{ds.O}$. In any case, the degenerated output resistance of the circuit is relatively large at approximately $R_{DEG.CO} r_{o.CO} g_{m.CO}$,

$$R_{OUT} = r_{o.CO} + R_{DEG.CO} + R_{DEG.CO}\, r_{o.CO}\, g_{m.CO} \approx R_{DEG.CO}\, r_{o.CO}\, g_{m.CO} \tag{3.80}$$

where $R_{DEG.CO}$ and R_{OUT} are $0.5r_\pi$ and $0.5\beta r_o$ for BJTs and r_{ds} and $r_{ds}{}^2 g_m$ for MOSFETs. Note R_{OUT} would be infinitely large if base-width and channel-length modulation effects (i.e., differences in v_{CE} and v_{DS}) were

nonexistent, yet R_{OUT} for the cascode circuit, whose purpose is to eliminate these effects, is finite. The reason for this finite value is that changes in v_{OUT} induce small variations in the v_{CE} or v_{DS} of the bottom output mirror device via the modulation effects of the cascode, thereby introducing v_{CE} or v_{DS} differences between the mirroring transistors.

Because cascode devices Q_{CIN} and Q_{CO} simply channel current into and out of a basic mirror, the low-frequency small-signal current gain of the circuit (i.e., $A_{I.LF}$) is roughly the same as that of the basic mirror, except with mitigated (but not completely eliminated) base-width and channel-length modulation effects. To fully comprehend all loading effects, however, source and loading resistances are included in the analysis and the current gain is defined as the translation of source current i_s to load current i_{Load}. As such, low-frequency current gain $A_{I.LF}$ is the product of its constituent small-signal translations: i_s to v_{in}, v_{in} to mirror's base voltage $v_{b.M}$, $v_{b.M}$ to mirror's output transconductor current $i_{gm.M}$, $i_{gm.M}$ to Q_{CO}'s emitter voltage $v_{e.CO}$, $v_{e.CO}$ to Q_{CO}'s transconductor current $i_{gm.CO}$, $i_{gm.CO}$ to v_{out}, and v_{out} to i_{Load} or

$$
\begin{aligned}
A_{I.LF} \equiv \frac{i_{Load}}{i_s} &= \left(\frac{v_{in}}{i_s}\right)\left(\frac{v_{b.M}}{v_{in}}\right)\left(\frac{i_{gm.M}}{v_{b.M}}\right)\left(\frac{v_{e.CO}}{i_{gm.M}}\right)\left(\frac{i_{gm.CO}}{v_{e.CO}}\right)\left(\frac{v_{out}}{i_{gm.CO}}\right)\left(\frac{i_{Load}}{v_{out}}\right) \\
&= (R_S \,\|\, R_{IN})\left[\frac{\left(\frac{1}{g_{m.M}}\right)}{R_{IN}}\right](-g_{m.M})\left[\frac{r_{o.CO} + R_{Load}}{1 + g_{m.CO} r_{o.CO}} \,\|\, \left(r_{\pi.CIN} + \frac{1}{g_{m.M}} + \frac{1}{g_{m.CIN}}\right)\right] \\
&\quad \times (g_{m.CO})(R_{OUT} \,\|\, R_{Load})\left(\frac{1}{R_{Load}}\right) \\
&\approx (R_S \,\|\, R_{IN})\left(\frac{1}{2}\right)(-g_{m.M})\left(\frac{1}{g_{m.CO}}\right)(g_{m.CO})\left(\frac{R_{OUT} \,\|\, R_{Load}}{R_{Load}}\right) \\
&\approx \left(\frac{2}{g_{m.M}}\right)\left(\frac{1}{2}\right)(-g_{m.M})\left(\frac{1}{g_{m.CO}}\right)(g_{m.CO})(1) \approx -1
\end{aligned}
\tag{3.81}
$$

As with any mirror, input and output resistances R_{IN} and R_{OUT} should approach zero and infinity, respectively, for maximum current gain.

Small-Signal Response at High Frequency

Given the basic nature of a current mirror, and even more so in a cascoded mirror, output pole p_O normally precedes input pole p_{IN} because R_{OUT} is considerably larger than R_{IN}. With this in mind, the parasitic capacitances present at v_{out} shunt R_{OUT} at and past output pole p_O. These capacitors include C_{Load} and Q_{CO}'s C_μ, except the latter has two $1/g_m$ resistances in series. The effects of these $1/g_m$ resistances,

however, are negligible because their collective resistance is substantially low with respect to $R_{OUT} \,||\, R_{Load}$, which means p_O reduces to $1/2\pi R_{Load}(C_{Load}+C_{\mu.CO})$:

$$\left(\frac{1}{sC_{\mu.CO}}+\frac{2}{g_m}\right)\Big\|\frac{1}{sC_{Load}} \approx \left.\frac{1}{s(C_{\mu.CO}+C_{Load})}\right|_{p_O \approx \frac{1}{2\pi R_{Load}(C_{\mu.CO}+C_{Load})}}$$

$$R_{OUT} \,||\, R_{Load} \approx (R_{DEG.CO} r_{o.CO} g_{m.CO}) \,||\, R_{Load} \approx R_{Load} \tag{3.82}$$

where $R_{EO.DEG}$ is $0.5r_\pi$ in the BJT case and r_{ds} in the MOS counterpart.

As in the basic mirror, but now across cascode transistor Q_{CO}, $C_{\mu.CO}$ feeds forward an out-of-phase signal from v_{in} to v_{out} with respect to mirror output current $i_{o.M}$. The resulting right-half-plane zero asserts its influence past roughly $g_{m.M}/2\pi C_{\mu.CO}$, when $C_{\mu.CO}$'s current equals the mirror's. The rest of the poles and zeros in the circuit are at similarly high frequencies because their respective resistances are low at approximately $1/g_m$ (i.e., Q_{CO}'s emitter resistance is also $1/g_m$ because C_{Load} is a short circuit at these frequencies and therefore exerts negligible loading effects on Q_{CO}'s emitter $1/g_m$ resistance). In traversing the circuit forward from input current source i_s, poles exist at v_{in}, the base of the basic mirror (i.e., $v_{b.M}$), Q_{CO}'s emitter, and v_{out}, as already defined by p_O. Given input resistance R_{IN} is substantially low, most of i_s flows through R_{IN} and not R_S, so the latter has little impact on frequency response. In the same vein, a negligible portion of i_s flows to Q_{CO}'s base because the resistance Q_{CO} presents (i.e., $Z_{IN.CO}$) is substantially high (since emitter degenerated Q_{CO} amplifies mirror output resistance $r_{o.M}$ and divides mirror output capacitance $C_{\mu.M}$ by $1 + \beta$).

Considering only input cascode Q_{CIN}'s $r_{\pi.CIN}$ and $C_{\pi.CIN}$ remain in the input path to the basic mirror (at $v_{b.M}$), most of i_s ultimately reaches $v_{b.M}$ because the effect of $C_{\pi.CIN}$ is also to channel whatever portion of i_s it carries from $v_{b.M}$ to the mirror. As such, input pole p_{IN} appears when the parasitic capacitance at $v_{b.M}$ shunts the $1/g_{m.M}$ resistance the basic mirror presents. Note the Miller effects on $C_{\mu.M}$ are low because the inverting gain across its terminals is low at roughly $g_{m.M}/g_{m.CO}$ or 1. In the end, the basic mirror's two C_π's and C_μ shunt $1/g_{m.M}$ past p_{IN} or $g_{m.M}/2\pi(2C_{\pi.M} + C_{\mu.M})$:

$$\left.\frac{1}{s(2C_{\pi.M}+C_{\mu.M})}\right|_{p_{IN} \approx \frac{g_{m.M}}{2\pi(2C_{\pi.M}+C_{\mu.M})}} \equiv \frac{1}{g_{m.M}} \tag{3.83}$$

Cascode transistor Q_{CO}'s emitter node $v_{e.CO}$ introduces what amounts to a closely spaced pole-zero pair because the equivalent capacitance at $v_{e.CO}$ at those frequencies has a series resistance that is only slightly larger than the total resistance present at $v_{e.CO}$. For one,

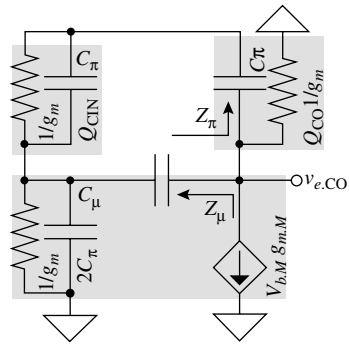


Figure 3.26 High-frequency small-signal equivalent circuit of the cascoded mirror with a moderately large source resistance (i.e., $R_s >> R_{IN}$) and load capacitance (i.e., output pole p_o is dominant).

the resistance at $v_{e.CO}$, as stated earlier, reduces to Q_{CO}'s $1/g_m$, as shown in Fig. 3.26. Secondly, the impedance through C_π is considerably lower than C_μ's (i.e., $Z_\pi << Z_\mu$, when referring to the figure) so C_μ's impact on $v_{e.CO}$ is lower. As a result, the mirroring transistor's voltage gain $v_{e.CO}/v_{b.M}$ is roughly $g_{m.M}/g_{m.CO}$ or 1 at low frequencies:

$$\frac{v_{e.CO}}{v_{b.M}} \approx g_{m.M}\left[\left(\frac{1}{g_{m.CO}}\right)\|\left(\frac{1}{sC_{\pi.CO}}+\frac{2}{g_m}\right)\|\left(\frac{1}{sC_{\mu.M}}+\frac{1}{g_m}\right)\right]$$

$$\approx g_{m.M}\left[\left(\frac{1}{g_{m.CO}}\right)\|\left(\frac{1}{sC_{\pi.CO}}+\frac{2}{g_m}\right)\right] \tag{3.84}$$

As frequency increases, Q_{CO}'s C_π shunts Q_{CO}'s $1/g_m$ and the gain begins to drop (in pole-like fashion). When C_π's impedance is sufficiently below the series $1/g_m$ resistances present, the gain flattens (i.e., a zero emerges), but because the impedance is not much lower than when it started, the zero is close to the pole. At and beyond these frequencies, though, all capacitors conduct displacement current and generally begin short-circuiting to the point frequency response no longer matters because the transistors are already past their respective transitional frequencies (i.e., f_T's).

3.3.3 Summary

Table 3.2 summarizes the performance parameters discussed in the basic and cascoded mirror subsections. With respect to large-signal response, the basic mirror offers the lowest minimum input voltage, but the cascode is able to match it by deriving a voltage bias for its cascode devices from another source. This fix, however, as with any other, implies tradeoffs, such as additional quiescent current and silicon real estate. Although the cascode circuit can decrease its output voltage headroom to two $V_{CE(min)}$'s or two $V_{DS(sat)}$'s by again deriving its bias from an off-line source, it cannot match the basic mirror's limit of one $V_{CE(min)}$ or one $V_{DS(sat)}$. Ultimately, the real benefits

	Basic Mirror w/R_β	Low-V. Cascoded Mirror
$V_{IN(min)}$	V_{BE} or V_{GS}	V_{BE} or V_{GS}
$V_{OUT(min)}$	$V_{CE(min)}$ or $V_{DS(sat)}$	$2V_{CE(min)}$ or $2V_{DS(sat)}$
Accuracy	V_A or λ error	–
R_{IN}	$1/g_m$	$1/g_m$
R_{OUT}	r_o or r_{ds}	$0.5r_\pi r_o g_m$ or $r_{ds}^2 g_m$
p_{IN}	$\frac{g_m}{2\pi(2C_\pi + C_\mu)}$ or $\frac{g_m}{2\pi(2C_{GS} + C_{GD})}$ (high)	
p_o	$\frac{1}{2\pi R_{Load}(C_\mu + C_{Load})}$ or $\frac{1}{2\pi R_{Load}(C_{GD} + C_{Load})}$ (low)	
z_{RHP}	$\frac{g_m}{2\pi C_\mu}$ or $\frac{g_m}{2\pi C_{GD}}$ (higher)	
Peculiarities		Extra pole-zero pair

TABLE 3.2 Mirror Summary

of the cascode circuit are lower base-width and channel-length modulation effects and higher output resistance because other important parameters such as input resistance, input pole, output pole, and right-half-plane zero are similar in both cases, except the cascode circuit introduces an additional pole-zero doublet. Note there are several ways of decreasing the adverse effects of β errors on accuracy in BJTs and the most benign in terms of tradeoffs and simplicity (but not the most accurate) is a series base resistor.

3.4 Five-Transistor Differential Amplifier

In the same spirit of combining single-transistor amplifiers and buffers to build differential input pairs and current mirrors, the latter circuits combine to produce slightly more complicated but rather useful circuits. Consider, for instance, the mirror-loaded differential pair shown in Fig. 3.27*a*, otherwise commonly known as the *five-transistor differential amplifier*. The purpose of the current mirror is to convert the differential output of the input pair into a single-ended signal without the current or gain loss associated with just using one resistor-loaded output terminal. With respect to large-signal response, mirror M_{PMP}-M_{PMN} superimposes and adds M_{NDP}'s output current variation (from zero to I_{Tail}), which would otherwise be unused in a single-ended output, to M_{NDN}'s (from $-I_{Tail}$ to zero). The combined currents, as a result,

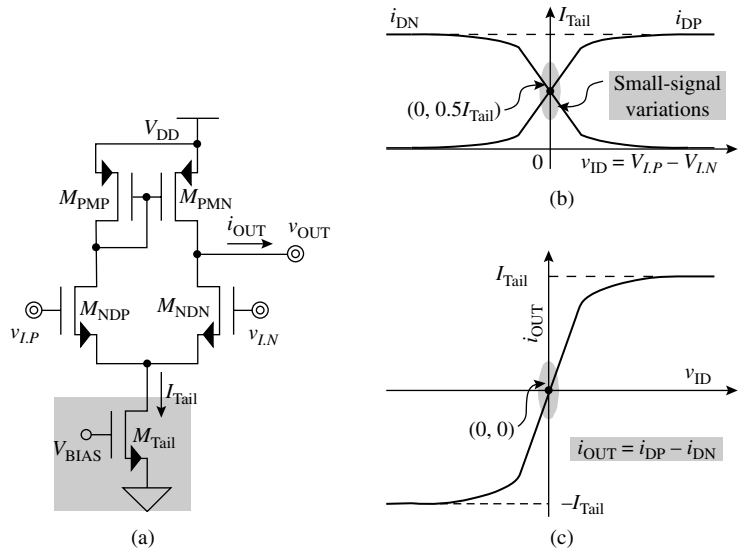


Figure 3.27 Mirror-loaded differential stage (*a*) circuit and (*b*) and (*c*) corresponding large-signal response.

yield a total output current variation Δi_{OUT} that is equivalent to twice its single-ended counterpart, which is to say i_{OUT} traverses from $-I_{Tail}$ to $+I_{Tail}$, as shown in Fig. 3.27*b*, instead of zero to $+I_{Tail}$ or $-I_{Tail}$ to zero. This doubling effect gives output v_{OUT} a *push-pull* quality because M_{PMN} can *push* as much current into v_{OUT} as M_{NDN} is able to *pull*.

3.4.1 Differential Signals

Similarly, with respect to small signals, mirror M_{PMP}-M_{PMN}, as a double-to-single signal converter, superimposes M_{NDP}'s positive differential current i_{dp} to M_{NDN}'s negative differential current i_{dn} to augment their collective gain to twice their single-ended counterpart. Evaluating the amplifier's small-signal equivalent circuit shown in Fig. 3.28*a* corroborates that the single-ended output is literally the differential output equivalent of input pair M_{NDP}-M_{NDN} because small-signal output v_{out} combines the effects of both small-signal currents $+i_{dp}$ and $-i_{dn}$ into the equivalent resistance present. As a result, the circuit produces a differential low-frequency gain $A_{D.LF}$ that is twice its single-ended equivalent:

$$\begin{aligned} A_{D.LF} \equiv \frac{v_{out}}{v_{id}} &= \frac{(i_{dp} - i_{dn})(r_{sd.PMN} \parallel r_{ds.NDN})}{v_{id}} \\ &= \frac{[0.5v_{id} - (-0.5v_{id})]g_{m.D}(r_{sd.PMN} \parallel r_{ds.NDN})}{v_{id}} \\ &= g_{m.D}(r_{sd.PMN} \parallel r_{ds.NDN}) \end{aligned} \tag{3.85}$$

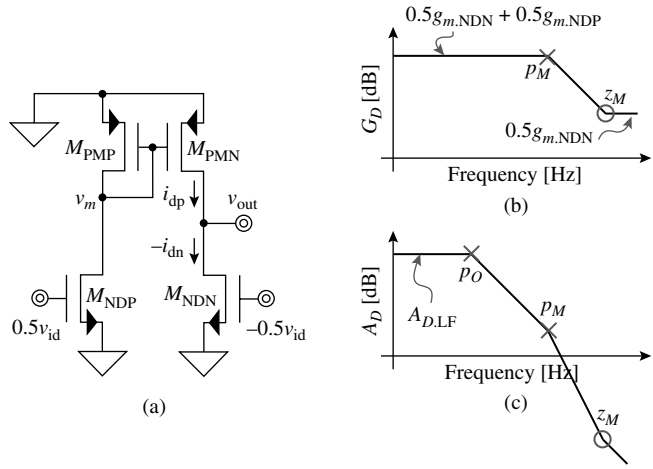


FIGURE 3.28 (*a*) Differential small-signal equivalent circuit of a mirror-loaded differential pair and its differential (*b*) transconductance G_D and (*c*) voltage gain A_D across frequency.

which means low-frequency differential transconductance $G_{D.LF}$ reduces to $g_{m.D}$, given the input pair comprises two matched and equally biased transistors:

$$G_{D.LF} = 0.5g_{m.NDP} + 0.5g_{m.NDN} = g_{m.D} \tag{3.86}$$

The frequency response of the mirror-loaded differential pair is the common-emitter/source amplifier associated with the output side of the input pair (i.e., M_{NDN}) and its complement (i.e., M_{NDP}) via a bandwidth-limited current-mirror translation to output v_{out}. To start, assuming the input pair's source resistance is low (i.e., there are no significant Miller effects in the input-pair transistors), the dominant pole of the circuit is at the output, at the highest resistance node. As such, the effects of output pole p_O become prevalent past the point the equivalent capacitance at v_{out} (i.e., C_{EQ}) shorts the equivalent resistance present, which is the parallel combination of M_{NDN}'s $r_{ds.NDN}$, M_{PMN}'s $r_{ds.PMN}$, and whatever loading resistance R_{Load} exists:

$$\frac{1}{s(C_{DG.PMN} + C_{DB.PMN} + C_{DG.NDN} + C_{DB.NDN} + C_L)} \equiv \frac{1}{sC_{EQ}}\bigg|_{p_O \simeq \frac{1}{2\pi(r_{sd.PMN} \| r_{ds.NDN} \| R_{Load})C_{EQ}}}$$

$$\equiv r_{sd.PMN} \| r_{ds.NDN} \| R_{Load} \tag{3.87}$$

Because the remaining effects in the differential pair (such as those of its out-of-phase feed-forward capacitors and their resulting right-half-plane zeros) and mirror are the result of resistances that are on the order of $1/g_m$, their effects are at higher frequencies. Of those, however, the parasitic capacitance associated with mirror node v_m is considerably larger (at $2C_{GS} + C_{GD}$) than the rest (at C_{GD}) so its pole effects appear first, after p_O. As a result, differential transconductance G_D, which is at its highest point at low frequencies (at $g_{m.D}$), starts dropping when the mirror's contribution decreases, past the point the total capacitance at the gates of mirror transistors M_{PMP} and M_{PMN} (i.e., $2C_{GS.M}$) shunts the equivalent resistance present (i.e., $1/g_{m.M}$), that is, past mirror pole p_M:

$$\left.\frac{1}{s(2C_{GS.M}+C_{GD.M})}\right|_{p_M \approx \frac{g_{m.M}}{2\pi(2C_{GS.M}+C_{GD.M})}} \equiv \frac{1}{g_{m.PMP}} \parallel r_{ds.PMP} \parallel r_{ds.NDP} \approx \frac{1}{g_{m.PMP}} \tag{3.88}$$

as shown in Fig. 3.28*b*.

At higher frequencies (past p_M), when mirror current i_{dp} falls to negligible levels, G_D reduces from $0.5g_{m.NDP} + 0.5g_{m.NDN}$ or $G_{D.LF}$ (which is differential transconductance $g_{m.D}$) to $0.5g_{m.NDN}$ or $0.5g_{m.D}$, half its low-frequency counterpart $G_{D.LF}$. The location of the left-half plane mirror zero (i.e., z_M) that flattens the dropping G_D to $0.5g_{m.D}$ corresponds to how much frequency must traverse to decrease low-frequency transconductance $G_{D.LF}$ (or $g_{m.D}$) to its target of $0.5G_{D.LF}$ (or $0.5g_{m.D}$), and because gain decreases linearly past pole p_M, z_M is $p_M\, g_{m.D}/0.5g_{m.D}$ or 2 times higher than p_M:

$$G_D = \left.\frac{G_{D.LF}}{\left(1+\frac{2\pi s}{p_M}\right)}\right|_{f>p_M} \approx \left.\frac{g_{m.D}}{\left(\frac{2\pi s}{p_M}\right)}\right\|_{z_M \approx 2p_M} \equiv 0.5G_{D.LF} = 0.5g_{m.D} \tag{3.89}$$

or algebraically,

$$G_D = \frac{0.5g_{m.NDP}}{\left(1+\frac{s}{p_M}\right)} + 0.5g_{m.NDN} = \frac{g_{m.NDP}\left(1+\frac{s}{2p_M}\right)}{\left(1+\frac{s}{p_M}\right)} \equiv \frac{g_{m.NDP}\left(1+\frac{s}{z_M}\right)}{\left(1+\frac{s}{p_M}\right)} \tag{3.90}$$

Because the pole and zero associated with the mirror are in close proximity, their effects tend to cancel. Nevertheless, p_M precedes z_M so

the phase dips slightly before z_M recovers it. This pole-zero doublet also appears in differential voltage gain A_D, as shown in Fig. 3.28*c*. Note the source resistance is not necessarily small and larger values pull input pole p_{IN} to lower frequencies.

3.4.2 Common-Mode Signals

The net effect of the mirror load, as appreciated in Fig. 3.29*a* and expected from the double-to-single signal conversion, is to offset the common-mode current in M_{NDN} (i.e., i_{cn}) with M_{NDP}'s (i.e., i_{cp}). Because the currents are approximately equal, their net effect is to produce an output voltage v_{out} and a common-mode gain A_C that near zero. The circuit, however, as it stands, does not ensure the dc drain-source voltages of the supposedly matched transistors in the mirror and differential pair equal, which means channel-length modulation causes a systematic offset in the currents (i.e., Δi) that ultimately produces a combined mirror error E_λ:

$$\Delta i = i_{cp} - i_{cn} = i_{cn}\left(\frac{i_{NDP}}{i_{cn}}\right)\left(\frac{i_{cp}}{i_{NDP}}\right) - i_{cn}$$

$$= i_{cn}\left[\left(\frac{1+\lambda V_{DS.NDP}}{1+\lambda V_{DS.NDN}}\right)\left(\frac{1+\lambda V_{SD.PMN}}{1+\lambda V_{SD.PMP}}\right) - 1\right] \equiv i_{cn}E_\lambda = v_{cm}G_{C.LF}E_\lambda \quad (3.91)$$

This offset in currents is proportional to common-mode input voltage v_{cm} and degenerated common-mode transconductance $G_{C.LF}$ and the

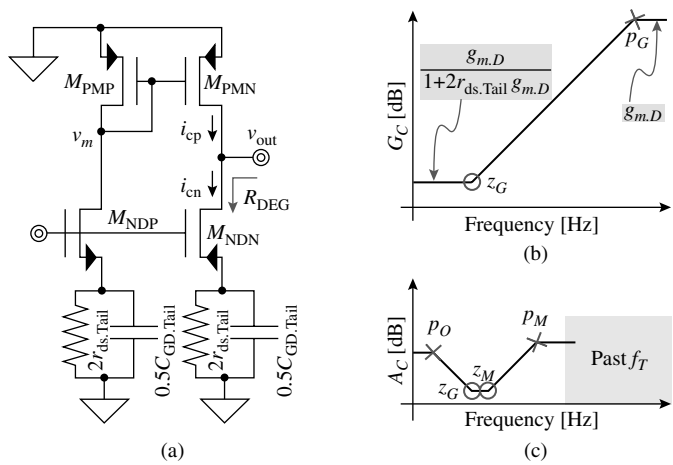


Figure 3.29 (*a*) Common-mode small-signal equivalent circuit of the mirror-loaded differential pair and its equivalent common-mode (*b*) transconductance G_c and (*c*) voltage gain A_c across frequency.

voltage it produces at v_{out} with respect to v_{cm} constitutes common-mode low-frequency gain $A_{C.LF}$:

$$A_{C.LF} \equiv \frac{v_{out}}{v_{cm}} = \frac{\Delta i_o r_{ds.PMN}}{v_{cm}} = \frac{(v_{cm} G_{C.LF} E_\lambda) r_{ds.PMN}}{v_{cm}}$$

$$= G_{C.LF} E_\lambda r_{ds.PMN} = \frac{g_{m.NDN} E_\lambda r_{ds.PMN}}{1 + g_{m.NDN} 2r_{ds.Tail}} \approx \frac{E_\lambda r_{ds.PMN}}{2r_{ds.Tail}} \qquad (3.92)$$

As a result, common-mode gain, which is a nondesirable effect of the circuit, is nonzero when mismatches in the circuit exist, and increases with increasing output resistance and degenerated transconductance values. Random and systematic mismatches in threshold voltages (i.e., V_T's), transconductance parameters (i.e., K''s), and channel-length modulation parameters (i.e., λ's) further increase offset error E_λ and degrade $A_{C.LF}$.

Assuming the input common-mode source resistance is low, the first pole to appear in the circuit corresponds to the node with the highest resistance, which in this case is normally v_{out}. The location of this pole is close to the location of the output pole in the differential circuit, except the differential pair is now degenerated and its output resistance (i.e., R_{DEG}) is considerably larger than the mirror's. Nonetheless, the equivalent capacitance at v_{out} shunts the equivalent resistance present past this output pole p_O:

$$\frac{1}{s(C_{DG.PMN} + C_{DB.PMN} + C_{DG.NDN} + C_{DB.NDN} + C_L)} \equiv \left.\frac{1}{sC_{EQ}}\right|_{p_O \approx \frac{1}{2\pi(r_{sd.PMN} \| R_{Load}) C_{EQ}}}$$

$$\equiv r_{sd.PMN} \| R_{Load} \| R_{DEG} \approx r_{sd.PMN} \| R_{Load} \qquad (3.93)$$

Degenerating resistance $2r_{ds.Tail}$ is also typically large and its effects may not be far from p_O, except shunting $2r_{ds.Tail}$ with $0.5C_{GD.Tail}$ increases common-mode transconductance G_C so its effect is to add zero z_G to A_C at approximately $1/2\pi r_{ds.Tail} C_{GD.Tail}$, as shown in Fig. 3.29*b*:

$$\left.\frac{1}{s(0.5C_{GD.Tail})}\right|_{z_G \approx \frac{1}{2\pi r_{ds.Tail} C_{GD.Tail}}} \equiv 2r_{ds.Tail} \qquad (3.94)$$

The feed-forward C_{GD} capacitors in the differential pair also tend to increase the value of the degenerated transconductance and shift z_G by increasing the net current available to the mirror. In any case, the pole that appears when degenerated transconductance G_C flattens to its nondegenerated state of $g_{m.D}$ (i.e., the input pair's transconductance)

is at considerably higher frequencies because the frequency that must traverse to increase degenerated low-frequency transconductance $G_{C.LF}$ or $1/2r_{ds.Tail}$ to $g_{m.D}$ is expansive:

$$G_C = G_{C.LF}\left(1+\frac{2\pi s}{z_G}\right)\Bigg|_{f>>z_G}$$

$$\approx \left(\frac{g_{m.D}}{1+2r_{ds.Tail}g_{m.D}}\right)\left(\frac{2\pi s}{z_G}\right)\Bigg|_{p_G\approx 2r_{ds.Tail}g_{m.D}z_G} \equiv g_{m.D} \tag{3.95}$$

where pole p_G is roughly $2r_{ds.Tail}g_{m.D}$ times higher than z_G. The current of the feed-forward capacitors overwhelms G_C at these frequencies so p_G becomes, for all practical purposes, inconsequential.

The effect of the mirror is peculiar because any small attenuation in compensating current i_{cp} represents an increase in offset current Δi and common-mode gain A_C. An increase in A_C amounts to a degradation in the circuit's ability to cancel i_{cn} with i_{cp}, the double-single signal conversion objective of the mirror. As a result, steering current away from the $1/g_{m.M}$ resistance at the input of the mirror with the parasitic capacitance present has an amplified impact on error E_λ (past pole p_M), which means the error increases at considerably lower frequencies with respect to p_M:

$$\frac{1}{s(2C_{GS.M}+C_{GD.M}+C_{GD.NDP})}\Bigg|_{p_M\approx\frac{g_{m.M}}{2\pi(2C_{GS.M}+C_{GD.M}+C_{GD.NDP})}} \equiv \frac{1}{g_{m.M}} \tag{3.96}$$

The zero z_M that results increases A_C at approximately E_λ times p_M, where E_λ is a small fraction (e.g., less than 4%) so z_M precedes p_M, as shown in Fig. 3.30:

$$\Delta i = i_{cp}-i_{cn} = \frac{i_{cn}(1+E_\lambda)}{\left(1+\frac{s}{p_M}\right)} - i_{cn} = \frac{i_{cn}(1+E_\lambda)-i_{cn}\left(1+\frac{s}{p_M}\right)}{\left(1+\frac{s}{p_M}\right)}$$

$$= \frac{i_{cn}E_\lambda\left(1+\frac{s}{p_ME_\lambda}\right)}{\left(1+\frac{s}{p_M}\right)} \equiv \frac{i_{cn}E_\lambda\left(1+\frac{s}{z_M}\right)}{\left(1+\frac{s}{p_M}\right)} \tag{3.97}$$

Note offset current Δi and the error peak when common-mode current i_{cp} is considerably smaller than its initial value of $i_{cn}(1 + E_\lambda)$, that is, when Δi nears i_{cn}, which happens at and past p_M.

In review, the only reason why there is a nonzero common-mode gain in the circuit is mismatches in the circuit exist, and its effects are

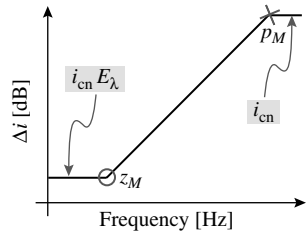


Figure 3.30 The effect of the load mirror on the differential pair's common-mode output offset current Δi.

less pronounced when degenerating resistance is highest. The loading capacitance present at v_{out}, the dominant portion of which is typically C_{Load}, tends to decrease common-mode gain A_C first, as shown in Fig. 3.29*c*. Tail current capacitance $C_{DG.Tail}$ increases transconductance G_C at around the same frequencies mirror capacitance $2C_{GS.M}$ increases offset current Δi, and feed-forward capacitors C_{GD} in the differential pair compound these effects near similar frequencies. Offset current Δi and the error it produces eventually peak when the mirror pole renders the mirror inactive past mirror pole p_M. Ultimately, because p_O and p_M in the common-mode circuit resemble those of the differential counterpart, their effects in common-mode rejection ratio CMRR are for all practical purposes inconsequential (because they cancel):

$$\text{CMRR} \equiv \frac{A_D}{A_C} \approx \frac{g_{m.D}(r_{sd.PMN} \parallel r_{ds.NDN})\left(1+\frac{2\pi s}{p_O}\right)\left(1+\frac{2\pi s}{p_M}\right)}{\left[\frac{g_{m.D}E_\lambda r_{sd.PMN}}{1+2r_{sd.Tail}g_{m.D}}\right]\left(1+\frac{2\pi s}{p_O}\right)\left(1+\frac{2\pi s}{z_G}\right)\left(1+\frac{2\pi s}{z_M}\right)\left(1+\frac{2\pi s}{p_M}\right)}$$

$$\approx \frac{g_{m.D}r_{sd.Tail}}{E_\lambda\left(1+\frac{2\pi s}{z_G}\right)\left(1+\frac{2\pi s}{z_M}\right)} \tag{3.98}$$

Differential gain A_D, however, does not match the transconductance and mirror zeros z_G and z_M in A_C so the zeros have a degrading effect on CMRR, as though they were poles.

3.5 Summary

Analog IC design, for the most part, decomposes into a series of single- and two-transistor circuits. Because transistors are, by definition, two-terminal resistors with a controlling third terminal, the manner in which a single transistor can be connected conforms to one of three basic configurations: common-emitter/source, common-collector/drain, and common-base/gate circuits. Common-emitter/source transistors amplify voltages and convert voltages to currents, and are therefore often useful in voltage and transconductance amplifiers.

Common-collector/drain followers are good voltage buffers with relatively low output impedances, which is why they are often useful in current-driving applications. Common-base/gate circuits buffer currents and therefore provide a means of channeling current into other current-processing circuits, including loading resistors, providing in the process a transimpedance gain.

With respect to small signals, the gate presents the highest resistance possible, followed by the collector or drain and then the base. The emitter or source terminal usually offers the least resistance at approximately $1/g_m$. The base-collector or gate-drain capacitor (with the help of the base-emitter or gate-source capacitor) not only steers energy away from its base or gate, having the effect of a low-frequency pole at the input, but also offers an out-of-phase, feed-forward path to the output, the result of which is a right-half-plane zero.

Two unique and valuable two-transistor circuits are differential pairs and current mirrors. The differential pair is useful in rejecting common-mode noise, which is especially prevalent and therefore problematic in mixed-signal systems sharing a common silicon substrate. Its ability to process differential signals also creates additional niche applications for the circuit. With respect to small signals, the differential pair ultimately decomposes into two common-emitter/source transistors. As to the current mirror, it offers a basic and robust means of processing currents, folding them so that sourcing a current into its input causes its output to sink a proportional fraction or gain of the same, and vice versa. As in the differential pair, the mirror decomposes into two common-emitter/source transistors, one of which is diode connected. Additional resistors and transistors may be used to improve its basic performance, as is the case in mitigating the effects of base-current errors and base-width and channel-length modulation on the output current. Ultimately, unique combinations of these basic one- and two-transistor analog building blocks comprise the foundation for more complex analog systems (as in the case of the five-transistor differential amplifier) performing a wide range of higher-order functions, the most pertinent of which for this text is regulation. Chapter 4 consequently reviews and discusses how negative feedback loops, which are nothing more than a combination of several signal-processing analog circuit blocks in a loop, regulate currents and voltages against variations in their operational and environmental conditions.

CHAPTER 4
Negative Feedback

Just as an artist applies the mechanics of language to poetry, an analog integrated circuit (IC) design engineer applies the basics of solid-state devices and circuit theory to microelectronic design. To that ultimate end, this chapter seeks to combine the devices and analog building blocks discussed in the previous two chapters to realize higher order functions, like voltage regulation, which is a driving force in this textbook. The fact is a voltage regulator must regulate and remain stable over a wide range of operating conditions. Negative-feedback circuits, in short, achieve the regulation features sought but in doing so they pose restrictions and generate by-products that may or may not be desirable in the system. In the end, the control limits and dynamics of the feedback loop determine the efficacy of the regulator with respect to ac, dc, and transient accuracy, the meets and bounds of the system. This chapter is therefore about negative feedback and follow-up chapters apply the fundamental teachings from this discussion to the design of linear regulators.

4.1 Generalities

4.1.1 Loop Composition

Feedback, in its most basic form, is a signal-processing *loop*, as illustrated in Fig. 4.1*a*. The loop is inconsequential on its own but ultimately useful as a signal-processing medium for incoming voltages, currents, time, and whatever signals come its way. Injection points *mix* these input signals (e.g., s_{I1} and s_{I2}) into the loop and extraction points *sense* or *sample* signals out of the loop for further processing or as stand-alone outputs (e.g., s_{O1} and s_{O2}).

From the perspective of processing input signal s_I into output s_O, the loop comprises a mixer, a forward open-loop gain A_{OL}, a sampler, and a feedback open-loop sense factor β_{FB}, as depicted in Fig. 4.1*b*. The loop signal with which s_I mixes, that is, feedback signal s_{FB}, must have the same dimensional unit as s_I—if one is a current, so must the other be, and so on with voltages and any other possible dimensional units. The absolute gain through the loop is the *loop gain* (LG), which

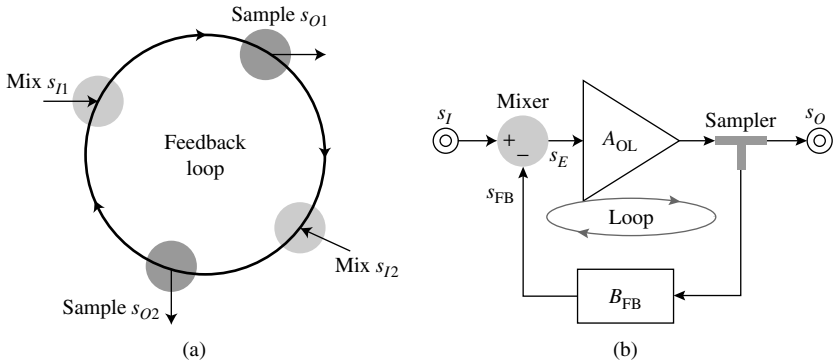


Figure 4.1 (*a*) Conceptual and (*b*) symbolic representations of a feedback loop.

is equal to $A_{OL}\beta_{FB}$. Because the loop establishes *negative* feedback, the gain through the loop, including the mixer, must, as a matter of course, be inverting (i.e., $-A_{OL}\beta_{FB}$). A negative gain, however, is not necessarily always the goal but the foregoing discussions will concentrate on negative feedback because of its vast application to analog circuits and, more specifically, regulators.

4.1.2 Gain

Ideally, the loop gain is infinitely high and can therefore only be stable when incoming signal s_I is equal to feedback signal s_{FB}, at which point error signal s_E is sufficiently low to produce a finite, quantifiable, and therefore stable output signal s_O. As a result, a feedback circuit, in qualitative terms, does what it can to force (or regulate) s_I to equal s_{FB} (which is equivalent to reducing s_E to a small fraction of s_I) and the extent of its success depends on how high loop gain LG (or $A_{OL}\beta_{FB}$) is in practice. The resulting closed-loop gain from s_I to s_{FB} is close to unity and the relationship between s_I and s_{FB} can be therefore described as a *virtual short* in the case of voltages and a *virtual mirror* in the case of currents. Similarly, in traversing the loop back from s_{FB} (now that it roughly equals s_I), s_E is only a translation of s_I through s_{FB}, which means s_E is s_I (or s_{FB}) divided by β_{FB} and again by A_{OL}:

$$s_E = s_I - s_{FB} = s_I - s_E(A_{OL}\beta_{FB}) = \frac{s_I}{1 + A_{OL}\beta_{FB}} \approx \frac{s_I}{A_{OL}\beta_{FB}} \tag{4.1}$$

or

$$s_{FB} = s_E(A_{OL}\beta_{FB}) = (s_I - s_{FB})(A_{OL}\beta_{FB}) = \frac{s_I(A_{OL}\beta_{FB})}{1 + A_{OL}\beta_{FB}} \approx s_I \tag{4.2}$$

assuming loop gain $A_{OL}\beta_{FB}$ is substantially larger than 1. As a result, as a signal-processing medium, because s_I and s_{FB} are virtually equal, output s_O is approximately only a β_{FB} translation of s_I (i.e., $s_O = s_{FB}/\beta_{FB} \approx s_I/\beta_{FB}$):

$$s_O = (s_I - s_{FB})A_{OL} = (s_I - s_O\beta_{FB})A_{OL} = \frac{s_I A_{OL}}{1 + \beta_{FB}A_{OL}} \approx \frac{s_I}{\beta_{FB}} \tag{4.3}$$

so closed-loop gain A_{CL} between s_I and s_O is roughly the reciprocal of feedback factor β_{FB}:

$$A_{CL} \equiv \frac{s_O}{s_I} = \frac{A_{OL}}{1 + A_{OL}\beta_{FB}} = A_{OL} \parallel \frac{1}{\beta_{FB}} \approx \frac{1}{\beta_{FB}} \tag{4.4}$$

Note negative feedback reduces A_{OL} to A_{CL} just as a $1/\beta_{FB}$ impedance would in the parallel combination of A_{OL} and $1/\beta_{FB}$. Viewing the effect of feedback on gain in this manner often helps infer other effects in the circuit.

As in the case of closed-loop gain A_{CL} with respect to forward open-loop gain A_{OL}, negative feedback amplifies or attenuates the effects of open-loop parameters by a factor of $1 + A_{OL}\beta_{FB}$. Consider, for instance, when A_{OL} is a first-order low-pass filter with a radians-per-second bandwidth of ω_{OL} and a low-frequency gain of $A_{OL.LF}$:

$$A_{OL} = \frac{A_{OL.LF}}{\left(1 + \dfrac{s}{\omega_{OL}}\right)} \propto R_{EQ} \parallel \frac{1}{sC_{EQ}} \tag{4.5}$$

Since closed-loop gain A_{CL} is the parallel impedance combination of A_{OL} and $1/\beta_{FB}$ and A_{OL} in this case is proportional to the parallel combination of its pole-setting resistance R_{EQ} and capacitance C_{EQ}, negative feedback shunts R_{EQ} with $1/\beta_{FB}$, thereby effectively reducing the pole-setting resistance to $R_{EQ} \parallel 1/\beta_{FB}$:

$$A_{CL} \propto \left(\frac{1}{\beta_{FB}} \parallel R_{EQ}\right) \parallel \frac{1}{sC_{EQ}} \tag{4.6}$$

which means the pole is now at considerably higher frequencies, assuming $1/\beta_{FB}$ is lower than R_{EQ}. In algebraic terms, substituting A_{OL} into the A_{CL} relationship yields a closed-loop corner frequency ω_{CL} that is $1 + A_{OL.LF}\beta_{FB}$ times larger than the original open-loop bandwidth:

$$A_{CL} = \frac{\left(\frac{A_{OL.LF}}{1+\frac{s}{\omega_{OL}}}\right)}{1+\left(\frac{A_{OL.LF}}{1+\frac{s}{\omega_{OL}}}\right)\beta_{FB}} = \frac{A_{OL.LF}}{1+\frac{s}{\omega_{OL}}+A_{OL.LF}\beta_{FB}}$$

$$= \frac{\left(\frac{A_{OL.LF}}{1+A_{OL.LF}\beta_{FB}}\right)}{\left(1+\frac{s}{\omega_{OL}(1+A_{OL.LF}\beta_{FB})}\right)} = \frac{A_{CL.LF}}{\left(1+\frac{s}{\omega_{CL}}\right)} \tag{4.7}$$

where $A_{CL.LF}$ is the low-frequency gain of the closed-loop system. Similarly, negative feedback reduces the corner frequency of a high-pass filter circuit ω_{OL} (i.e., extends its bandwidth) by a factor of $1 + A_{OL.LF}\beta_{FB}$:

$$A_{OL} = \frac{A_{OL.LF}\left(\frac{s}{\omega_{OL}}\right)}{\left(1+\frac{s}{\omega_{OL}}\right)} \tag{4.8}$$

or

$$A_{CL} = \frac{\left[\frac{A_{OL.LF}\left(\frac{s}{\omega_{OL}}\right)}{1+\frac{s}{\omega_{OL}}}\right]}{1+\left[\frac{A_{OL.LF}\left(\frac{s}{\omega_{OL}}\right)}{1+\frac{s}{\omega_{OL}}}\right]\beta_{FB}} = \frac{A_{OL.LF}\left(\frac{s}{\omega_{OL}}\right)}{1+\left(\frac{s}{\omega_{OL}}\right)+A_{OL.LF}\beta_{FB}\left(\frac{s}{\omega_{OL}}\right)}$$

$$= \frac{A_{OL.LF}\left(\frac{s}{\omega_{OL}}\right)}{\left[1+\frac{s(1+A_{OL.LF}\beta_{FB})}{\omega_{OL}}\right]}$$

$$= \frac{\left(\frac{A_{OL.LF}}{1+A_{OL.LF}\beta_{FB}}\right)\left[\frac{s(1+A_{OL.LF}\beta_{FB})}{\omega_{OL}}\right]}{\left[1+\frac{s(1+A_{OL.LF}\beta_{FB})}{\omega_{OL}}\right]} = \frac{A_{CL.LF}\left(\frac{s}{\omega_{CL}}\right)}{\left(1+\frac{s}{\omega_{CL}}\right)} \tag{4.9}$$

As with gain and bandwidth, negative feedback also reduces (i.e., improves) the sensitivity of an amplifier by a $1 + A_{OL}\beta_{FB}$ factor. Sensitivity refers to how vulnerable a parameter is to external forces, in other words, its percentage change with respect to its ideal value in response to external variations:

$$S_{A_{OL}} \equiv \frac{dA_{OL}}{A_{OL}} \tag{4.10}$$

where $S_{A_{OL}}$ is the sensitivity of A_{OL} and dA_{OL} the variation in A_{OL} that results from a change of some external force, which is another way of saying its first derivative with respect to that force. Qualitatively, because A_{CL} is approximately $1/\beta_{FB}$ and roughly independent of A_{OL}, A_{CL}'s sensitivity to A_{OL} is considerably low. Algebraically, applying the same sensitivity concept to A_{CL} as done to A_{OL}, but using its relationship with respect to A_{OL}, reveals closed-loop sensitivity $S_{A_{CL}}$ is lower than open-loop sensitivity $S_{A_{OL}}$ by a factor of $1 + A_{OL}\beta_{FB}$:

$$S_{A_{CL}} \equiv \frac{dA_{CL}}{A_{CL}} = \frac{d\left(\frac{A_{OL}}{1+A_{OL}\beta_{FB}}\right)}{\left(\frac{A_{OL}}{1+A_{OL}\beta_{FB}}\right)} = \frac{\left(\frac{(1+A_{OL}\beta_{FB})dA_{OL} - (A_{OL})\beta_{FB}\,dA_{OL}}{(1+A_{OL}\beta_{FB})^2}\right)}{\left(\frac{A_{OL}}{1+A_{OL}\beta_{FB}}\right)}$$

$$= \frac{dA_{OL}}{A_{OL}(1+A_{OL}\beta_{FB})} = \frac{S_{A_{OL}}}{1+A_{OL}\beta_{FB}} \tag{4.11}$$

Along the same lines, because A_{CL} is roughly independent of A_{OL}, negative feedback reduces distortion that results from variations in A_{OL} across a signal's amplitude; in other words, it linearizes a circuit. Similarly, just as negative feedback reduces forward open-loop gain, extends bandwidth, desensitizes a circuit, and linearizes its response by a factor of $1 + A_{OL}\beta_{FB}$, negative feedback also increases or decreases open-loop impedances by a factor of $1 + A_{OL}\beta_{FB}$, as will be seen in follow-up discussions.

4.2 Mixers

The mixer determines several characteristics of the loop, from the signal relationships across the loop (e.g., s_I sets s_{FB} and s_{FB} in turn sets s_O, s_E, and all other loop signals) to the circuit's input impedance, and its results depend on the manner in which the circuit mixes the incoming signal into the loop. A mixing function, to start, is nothing more than a summer whose output is the difference of its inputs (e.g., s_E is the difference of s_I and s_{FB} in Fig. 4.1*b*). In the first case, adding or

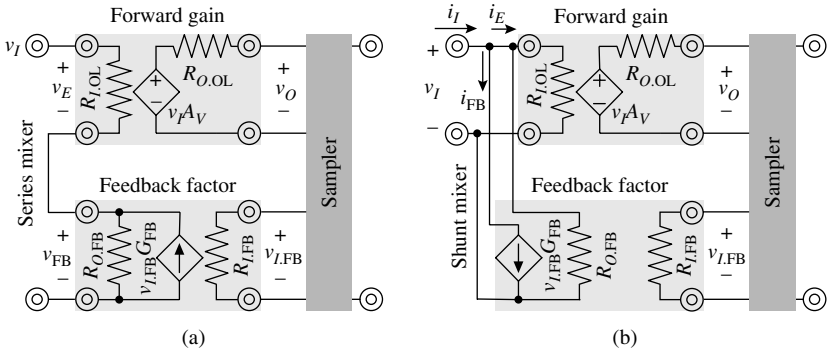


Figure 4.2 (*a*) Series- and (*b*) shunt-mixed negative-feedback circuits.

subtracting voltages amounts to placing them in *series*, as shown in Fig. 4.2*a*, where two-port Norton-equivalent circuits model forward gain A_{OL} and feedback factor β_{FB}. Because A_{OL}'s input resistance $R_{I.OL}$ is in series with feedback factor β_{FB}, series feedback has a tendency to increase the effective input resistance of the forward open-loop amplifier (i.e., A_{OL} circuit). A parallel or shunt configuration, on the other hand, adds or subtracts currents, as illustrated in Fig. 4.2*b*, giving rise to the term *shunt* mixing. In this latter case, because $R_{I.OL}$ is now in parallel with feedback factor β_{FB}, negative feedback tends to decrease the input resistance of the A_{OL} amplifier. Since the effects of the mixer differ vastly from both series and shunt configurations, determining the nature of the mixing function is key in assessing its impact on input impedance.

Ascertaining the type of mixer being used amounts to locating the mixing agent, which can only reside at signal-injection points. Such a loop may have several mixers, but only the mixer corresponding to a particular input and the gain through the loop (i.e., LG or $A_{OL}\beta_{FB}$) determine the electrical characteristics of that terminal. As such, after identifying the loop as a negative feedback circuit, tracing the input signal into the loop is the second step in identifying the mixer for that input terminal. The next and final step is to recognize all possible mixers in the input path and verify which one sums an incoming signal (i.e., s_I) with a feedback loop signal (i.e., s_{FB}) and pushes their difference ($|s_I - s_{FB}|$) on through the loop.

4.2.1 Series (Voltage) Mixers

The most commonly recognized voltage mixer is probably the differential voltage amplifier, otherwise known as the operational amplifier or *op amp*, for short (Fig. 4.3), because its output is (by definition) the amplified difference of its input voltages. As such, whenever an input signal finds its way into one of the input terminals of an op amp,

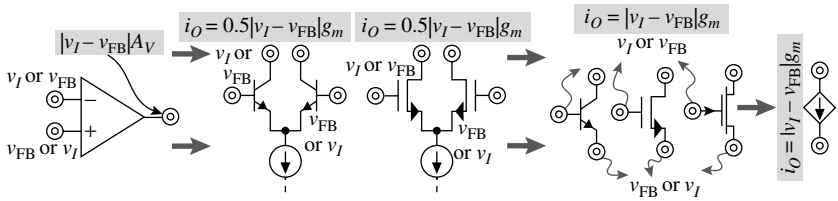


FIGURE 4.3 Descending abstraction levels of most common series (i.e., voltage) mixers: op amp, differential pairs, transistors, and difference transconductors.

wherein the other input terminal is in the ac signal path of the loop, the op amp is a voltage mixer. However, if one of the input terminals is at ac ground, that is, not in the feedback path, as in the case of an inverting op-amp configuration, the op amp is no longer a voltage mixer because its output is the amplified version of its *one* ac input (i.e., $v_O = (v_I - 0)A_V = v_i A_V$). In any case, the op amp is only an abstraction of a more primitive circuit, the differential pair, whose output current is directly proportional to the differential voltage across its inputs (Fig. 4.3). Even more fundamentally, the differential amplifier is a pair of common-emitter/source transistors whose individual transconductor currents are proportional to the ac differential voltage across their base-emitter/gate-source terminals (Fig. 4.3). As a result, if an input signal finds its way into a base, gate, emitter, or source and the complementary emitter-, source-, base-, or gate-input terminal is in the ac feedback path, the transistor's transconductor constitutes a voltage mixer.

The input resistance of a series mixer is the series combination of the output resistance of feedback factor β_{FB} (i.e., $R_{O.FB}$) and an amplified version of the open-loop input resistance of forward amplifier A_V (i.e., $R_{I.OL}$). Generally, referring to Fig. 4.2*a*, closed-loop input resistance $R_{I.CL}$ is the ratio of the ac voltage across its input terminals (i.e., v_i) and the induced small-signal current through the terminals (i.e., i_i). Input voltage v_i is the collective voltage across $R_{I.OL}$ and $R_{O.FB}$ and, assuming the output is a voltage, feedback factor β_{FB} is v_{FB}/v_O or the product of Norton-equivalent feedback transconductance G_{FB} and $R_{O.FB}$ (i.e., $\beta_{FB} = G_{FB}R_{O.FB}$) and forward open-loop gain $A_{V.OL}$ is v_o/v_e or A_V:

$$R_{I.CL} \equiv \frac{v_i}{i_i} = \frac{i_i R_{I.OL} + (i_i + i_{G.FB})R_{O.FB}}{i_i}$$

$$= \frac{i_i R_{I.OL} + [i_i + (i_i R_{I.OL})A_V G_{FB}]R_{O.FB}}{i_i}$$

$$= R_{I.OL}(1 + A_{OL}\beta_{FB}) + R_{O.FB} \qquad (4.12)$$

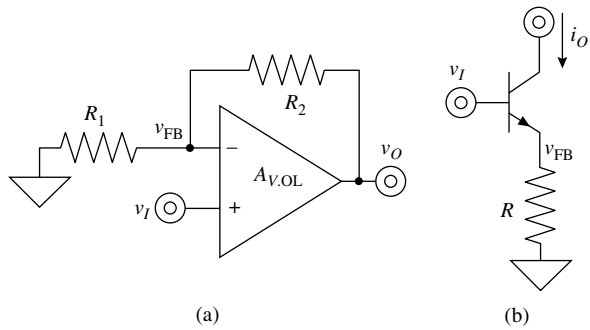


Figure 4.4 Series- (or voltage-) mixed examples: (*a*) noninverting op-amp configuration and (*b*) emitter-degenerated common-emitter (CE) transistor amplifier.

In other words, series-mixing increases open-loop input resistance $R_{I.OL}$ (and equivalent input impedance $Z_{I.OL}$) by a factor of $1 + A_{OL}\beta_{FB}$. Note that circuits with substantially increased input impedances constitute *ideal loads for incoming voltages* because they do not shunt input energy (i.e., signals) to ground.

The noninverting op-amp configuration shown in Fig. 4.4*a* is an example of a series-mixed negative-feedback circuit. The negative-feedback loop traverses from the inverting input, across the op amp, and through feedback resistor R_2 back to the inverting input. Because feedback signal v_{FB} is on the inverting input and input v_I on the complementary input terminal, the op amp series-mixes v_I into the feedback loop (with v_{FB}). Forward open-loop gain $A_{V.OL}$ is the transfer gain from the difference of its input terminals to output signal v_O (i.e., $A_{V.OL}$ is $v_o/|v_I - v_{FB}|$), which in this case is the differential gain of the op amp. Feedback factor β_{FB} is the transfer function from v_O back to the feedback input of the mixer (i.e., v_{FB}), which is the voltage divider ratio between resistors R_2 and R_1, that is, β_{FB} is $R_1/(R_1 + R_2)$. As a result, because v_O is only a β_{FB} translation of v_{FB}, which is virtually shorted to v_I, v_O is roughly v_I/β_{FB} or the well-known closed-loop gain of $(R_1 + R_2)/R_1$, assuming loop gain $A_{V.OL}\beta_{FB}$ is considerably larger than 1:

$$A_{V.CL} \equiv \frac{v_O}{v_I} = \frac{A_{V.OL}}{1 + A_{V.OL}\beta_{FB}} = \frac{A_{V.OL}}{1 + \left(\dfrac{A_{V.OL}R_1}{R_1 + R_2}\right)} \approx \frac{R_1 + R_2}{R_1} \qquad (4.13)$$

The closed-loop input resistance (i.e., $R_{I.CL}$) of the circuit is the series combination of the output resistance of the β_{FB} network $R_{O.FB}$ (which is the parallel combination of R_1 and R_2 in this example) and the amplified version of differential open-loop input resistance $R_{I.OL}$, which in the case of a BJT differential input pair is an amplified translation of $2r_\pi$:

$$R_{I.CL} = R_{O.FB} + (1 + A_{V.OL}\beta_{FB})R_{I.OL}$$
$$= (R_1 \parallel R_2) + \left(1 + \frac{A_{V.OL}R_1}{R_1 + R_2}\right)(2r_\pi) \tag{4.14}$$

where r_π approaches infinity in the MOS case.

Perhaps more fundamental series-mixed negative-feedback circuits are degenerated common-emitter (CE) and common-source (CS) amplifiers (Fig. 4.4*b*). In such circuits, the transistor and its degenerating resistor comprise a negative feedback loop because the transistor-resistor combination respond to oppose variations in output current i_O: as i_O increases, so does emitter/source voltage v_{FB}, causing the base-emitter or gate-source voltage to decrease and therefore producing a proportionally smaller transconductor current i_{gm} and i_O. Because input voltage v_I is at the base or gate and feedback loop signal v_{FB} at the emitter or source, the transistor's transconductor g_m mixes v_I with v_{FB} (i.e., transconductor current i_{gm} is the product of $|v_I - v_{FB}|$ and g_m). Since the output is a current and the input a voltage, the forward open-loop transconductance gain $A_{G.OL}$ and feedback factor β_{FB} are the transfer functions from $(v_I - v_{FB})$ to i_O (i.e., g_m) and i_O to v_{FB} (i.e., R), respectively. The resulting closed-loop transconductance gain $A_{G.CL}$ is therefore a loop-gain-*degenerated* version of g_m,

$$A_{G.CL} \equiv \frac{i_O}{v_I} = \frac{A_{G.OL}}{1 + A_{G.OL}\beta_{FB}} = \frac{g_m}{1 + g_m R} \approx \frac{1}{R} \tag{4.15}$$

and the closed-loop input resistance of the circuit (i.e., $R_{I.CL}$) is the series combination of R and the loop-gain translation of r_π or

$$R_{I.CL} = R_{O.FB} + (1 + A_{G.OL}\beta_{FB})R_{I.OL} \approx R + (1 + g_m R)r_\pi$$
$$= r_\pi + (1 + g_m r_\pi)R = r_\pi + (1 + \beta)R \tag{4.16}$$

As expected, the results do not differ from the ones obtained by conventional means, as discussed in Chap. 3. Note, however, β_{FB} in $R_{I.CL}$ for the BJT case includes an approximation because not all the emitter current flows through the collector, since the base drives a fraction; in other words, β_{FB} is

$$\beta_{FB} \equiv \frac{v_E}{i_C} = \left(\frac{i_E}{i_C}\right)R = \left(\frac{\beta_{BJT} + 1}{\beta_{BJT}}\right)R \approx R \tag{4.17}$$

assuming β_{BJT}, as the BJT's current gain, is considerably high (e.g., 50–100 A/A).

The transistor, because it also processes the *small-signal* differential voltage across its base-emitter or gate-source terminals, is equivalent to

a differential op amp with an output resistance equal to r_o or r_{ds}. The only differences are (1) a true op amp also processes the *large-signal* differential voltage across its inputs (i.e., dc voltage across op-amp inputs is zero and dc voltage across base-emitter or gate-source terminals is nonzero) and (2) the emitter or source terminal conducts current (while neither op-amp input does). In other words, virtual-short approximations apply to small signals in both the op amp and the transistor in negative feedback (i.e., $v_b \approx v_e$, $v_g \approx v_s$, and $v_+ \approx v_-$) but only to the op amp when considering large signals (i.e., $V_B \neq V_E$ and $V_G \neq V_S$ but $V_+ \approx V_-$). Given this result, Fig. 4.5*a* illustrates a transistor-level embodiment of the more general series-mixed op-amp case. Applying the above-stated approximation to this circuit indicates a virtual short exists between the base input and the corresponding emitter feedback terminal (i.e., $v_i \approx v_{fb}$) so the small-signal closed-loop output voltage is an amplified voltage-divider ratio of v_i (i.e., v_o is roughly v_i/β_{FB} or v_{fb}/β_{FB}):

$$A_{V.CL} \equiv \frac{v_o}{v_i} \approx \frac{v_o}{v_{fb}} = \frac{\left(\dfrac{v_{fb}}{R_1 \parallel R_{E1}}\right)[(R_1 \parallel R_{E1}) + R_2]}{v_{fb}} = \frac{(R_1 \parallel R_{E1}) + R_2}{(R_1 \parallel R_{E1})} \quad (4.18)$$

which is the transistor-level parallel of the noninverting op-amp configuration.

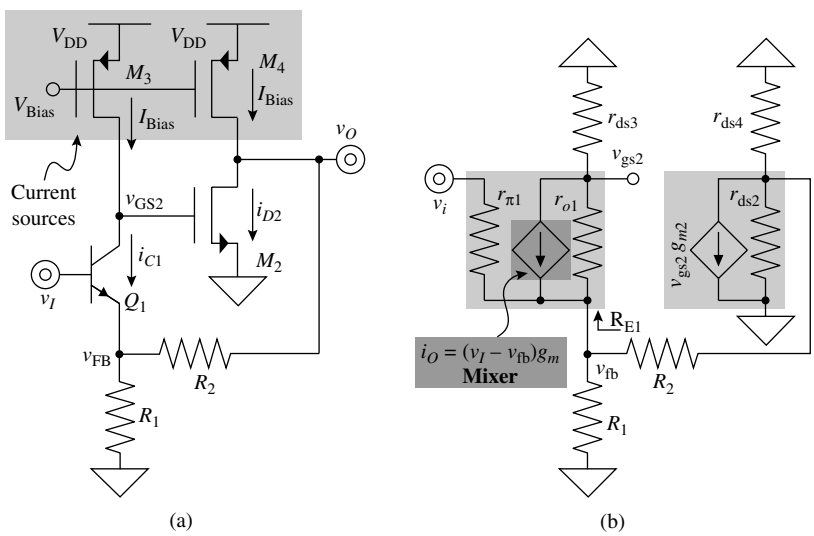


Figure 4.5 (*a*) A transistor-level embodiment of the series- (voltage-) mixed op-amp circuit in a noninverting configuration and (*b*) its corresponding low-frequency small-signal equivalent circuit.

The circuit shown in Fig. 4.5*a* is also an extrapolation of the degenerated circuit in Fig. 4.4*b* because additional circuits (like M_2) further process the transconductor current of the mixing transistor (i.e., Q_1) before feeding the signal back to the mixer's feedback point (e.g., Q_1's emitter). In other words, a negative feedback loop exists because Q_1 and M_2 and R_1 and R_2 respond to oppose variations in Q_1's collector current i_{C1}—an increase in i_{C1} causes v_{GS2} to decrease, i_{D2} to decrease, v_O (or v_{D2}) to increase, and v_{FB} (or v_{E1}) to increase, producing a lower v_{BE1} and consequently a proportionally smaller transconductor current i_{gm1} and i_{C1}. What is interesting about this case, and most larger loops employing series mixing, is the existence of two intertwined negative-feedback loops, as increases in i_{C1} also cause increases in v_{FB} directly, without going through the loop (i.e., degenerated transistor Q_1 is itself a negative feedback loop).

In multiple-loop cases, the loop gain of one loop normally overwhelms the others, ultimately determining the closed-loop dynamics of the circuit. For instance, the forward low-frequency small-signal open-loop voltage gain of the larger loop in Fig. 4.5*b* is the product of the gains from $(v_i - v_{fb})$ to v_{gs2} (which is roughly $-g_{m1}r_{ds3}$) and v_{gs2} to v_o (which is approximately $-g_{m2}\{r_{ds2} \| r_{ds4} \| [R_2 + (R_1 \| R_{E1})]\}$) so $A_{V.OL}$ is $g_{m1}r_{ds3}g_{m2}\{r_{ds2} \| r_{ds4} \| [R_2 + (R_1 \| R_E)]\}$. In opening the loop to determine R_{E1}, to avoid opening the two loops, since only one feedback loop is being considered, only v_{gs2} in M_2's transconductor *source* is shorted to zero (not the actual gate voltage) so emitter resistance R_E in the presence of M_3's load reduces to roughly $2/g_{m1}$:

$$R_{E1} = \left(\frac{r_{ds3} + r_{o1}}{g_{m1}r_{01}}\right) \| r_{\pi 1} \approx \frac{2}{g_{m1}} \tag{4.19}$$

Arbitrarily grounding Q_1's emitter would have otherwise masked not only R_1 but also Q_1's inherent feedback loop and similarly grounding M_2's gate would have concealed the effect of r_{ds3}. The circuit's open-loop feedback factor β_{FB} is the transfer function from v_o to v_{fb}, which is the voltage divider between R_2 and $R_1 \| R_E$ so β_{FB} is $(R_1 \| R_{E1})/[(R_1 \| R_{E1}) + R_2]$. The resulting loop gain can be substantially larger than that of the simple degenerated case, that is, larger than $g_{m1}(R_1 \| R_2)$,

$$\begin{aligned} \mathrm{LG} &\equiv A_{V.OL}\beta_{FB} \\ &\approx \left(g_{m1}r_{ds3}g_{m2} \left\{ r_{ds2} \| r_{ds4} \| \left[\left(R_1 \| \frac{2}{g_{m1}} \right) + R_2 \right] \right\} \right) \left[\frac{R_1 \| \frac{2}{g_{m1}}}{\left(R_1 \| \frac{2}{g_{m1}} \right) + R_2} \right] \end{aligned} \tag{4.20}$$

which is why the overall closed-loop voltage gain of the circuit approximates to

$$A_{V.\mathrm{CL}} = \frac{A_{V.\mathrm{OL}}}{1 + A_{V.\mathrm{OL}}\beta_{\mathrm{FB}}} \approx \frac{1}{\beta_{\mathrm{FB}}} = \frac{R_2 + \left(R_1 \,||\, \dfrac{2}{g_{m1}}\right)}{R_1 \,||\, \dfrac{2}{g_{m1}}} \tag{4.21}$$

Including the loading effects of a particular resistor on both intertwined loops may lead to redundancy and therefore produce inaccurate results. For instance, feedback factor's output resistance $R_{O.\mathrm{FB}}$, which in this case is $R_1 \,||\, R_2$, degenerates Q_1 and sets the open-loop input resistance of the outer loop (i.e., $R_{I.\mathrm{OL}}$) to $r_{\pi 1}+(1+\beta)(R_1 \,||\, R_2)$. Including $R_{O.\mathrm{FB}}$ as $R_1 \,||\, R_2$ again in the expression for closed-loop input resistance $R_{I.\mathrm{CL}}$ would incorrectly recount its effect on the circuit. The point is $R_{I.\mathrm{OL}}$ absorbs $R_{O.\mathrm{FB}}$ so $R_{O.\mathrm{FB}}$ disappears from $R_{I.\mathrm{CL}}$'s expression:

$$R_{I.\mathrm{CL}} = R_{I.\mathrm{OL}}(1 + A_{V.\mathrm{OL}}\beta_{\mathrm{FB}}) + R_{O.\mathrm{FB}} \approx R_{I.\mathrm{OL}}(1 + A_{V.\mathrm{OL}}\beta_{\mathrm{FB}}) \tag{4.22}$$

Note these results represent three sets of approximations: one loop overwhelms the other intertwined loop, the larger loop's gain is considerably larger than the smaller loop's, and general transistor-level impedance and small-signal gains linearize (i.e., approximate) the effects of a nonlinear circuit.

4.2.2 Shunt (Current) Mixers

In its most basic form, a current mixer is a star connection of input, feedback, and error currents, as illustrated with i_I, i_{FB}, and i_E in Fig. 4.2*b*. The output resistance of the feedback factor β_{FB} (i.e., $R_{O.\mathrm{FB}}$) is in parallel with the input resistance of the forward gain A_{OL} (i.e., $R_{I.\mathrm{OL}}$) and both of them combined constitute the effective open-loop input resistance of the circuit (i.e., $R_{O.\mathrm{FB}} \,||\, R_{I.\mathrm{OL}}$). Generally, finding this type of mixer in large feedback circuits is not always straightforward because applying input voltages to the circuit does not necessarily imply the feedback loop mixes voltages, as intuition would initially suggest. Even so, testing for series-mixed inputs by searching for base-emitter and/or gate-source terminals carrying input and feedback voltages, as discussed in the previous subsection, is easier and therefore recommended. After discarding the possibility of a series-mixed network, specifically looking for star-feedback connections helps identify probable shunt (current) mixers. Note the shunt-mixer approximation is a virtual mirror between incoming current i_I and feedback current i_{FB} because error current i_E is nearly zero.

The closed-loop input resistance of a shunt-mixed negative feedback circuit is the parallel combination of the effective open-loop

input resistance of the circuit (i.e., the parallel combination of $R_{I.OL}$ and $R_{O.FB}$) and its loop-gain-reduced counterpart. Referring to Fig. 4.2*b* and assuming the output is a voltage, β_{FB} is ratio i_{FB}/v_O or simply G_{FB} (i.e., β_{FB} is G_{FB}) and open-loop transimpedance gain $A_{R.OL}$ (i.e., A_{OL}) is ratio v_O/i_E or $v_O/(i_I\text{-}i_{FB})$ or $(R_{I.OL} \,||\, R_{O.FB})A_V$, where A_V is the voltage gain portion across the forward open-loop amplifier. Because any variation in voltage causes a change in i_I (through $R_{I.OL} \,||\, R_{O.FB}$) and the feedback loop responds to oppose that by sinking or sourcing a compensating feedback current i_{FB} (via G_{FB}), there is little voltage variation across the input. As a result, the resistance into the feedback transconductor G_{FB} (i.e., $R_{G.FB}$) is low, as derived from applying a test voltage (i.e., v_T) and dividing it by the resulting test current (i.e., i_T),

$$R_{G.FB} \equiv \frac{v_T}{i_T} = \frac{v_i}{i_{G.FB}} = \frac{i_E(R_{I.OL} \,||\, R_{O.FB})}{i_E A_{OL}\beta_{FB}} = \frac{R_{I.OL} \,||\, R_{O.FB}}{A_{OL}\beta_{FB}} \tag{4.23}$$

as is the resulting closed-loop input resistance $R_{I.CL}$, which is the parallel combination of $R_{G.FB}$, $R_{O.FB}$, and $R_{I.OL}$,

$$R_{I.CL} = (R_{O.FB} \,||\, R_{I.OL}) \,||\, \left(\frac{R_{O.FB} \,||\, R_{I.OL}}{A_{OL}\beta_{FB}}\right) = \frac{R_{O.FB} \,||\, R_{I.OL}}{1 + A_{OL}\beta_{FB}} \tag{4.24}$$

As a result, shunt mixing reduces the effective input resistance of the open-loop circuit by a factor of $1 + A_{OL}\beta_{FB}$, the same factor included in the expressions of most other closed-loop feedback parameters with respect to their open-loop counterparts. Since the input resistance (and impedance) can be substantially low and infinitesimally low input impedances are *ideal loads for incoming currents*, shunt-mixed circuits present desirable loads for incoming currents.

The inverting op-amp configuration shown in Fig. 4.6*a* is a shunt-mixed negative feedback circuit. The op amp is not a series mixer because its noninverting input is at ac ground and the star connection at its inverting terminal mixes the current flowing through resistor R_1 and feedback current i_{FB}. In spite the input to the overall circuit is a voltage, the current the voltage induces is what is being mixed and processed into the loop, making the feedback circuit a transimpedance amplifier. Before applying negative feedback theory, it is often useful to use standard negative-feedback (and op-amp) approximations, which in this case amount to a virtual short across the input terminals and a virtual mirror between i_I and feedback current i_{FB} or i_{R2} (i.e., the current flowing through R_2) so i_I equals i_{R2}. As such, the inverting terminal is virtual ground, i_I is v_{IN}/R_1, and v_O is the ohmic drop across R_2 or $-i_I R_2$ or $-v_{IN}R_2/R_1$, which translates to an overall voltage gain A_V (or v_O/v_{IN}) of $-R_2/R_1$.

Since all impedances affect the circuit, extracting dependent mixing and feedback current and/or voltage sources from two-port

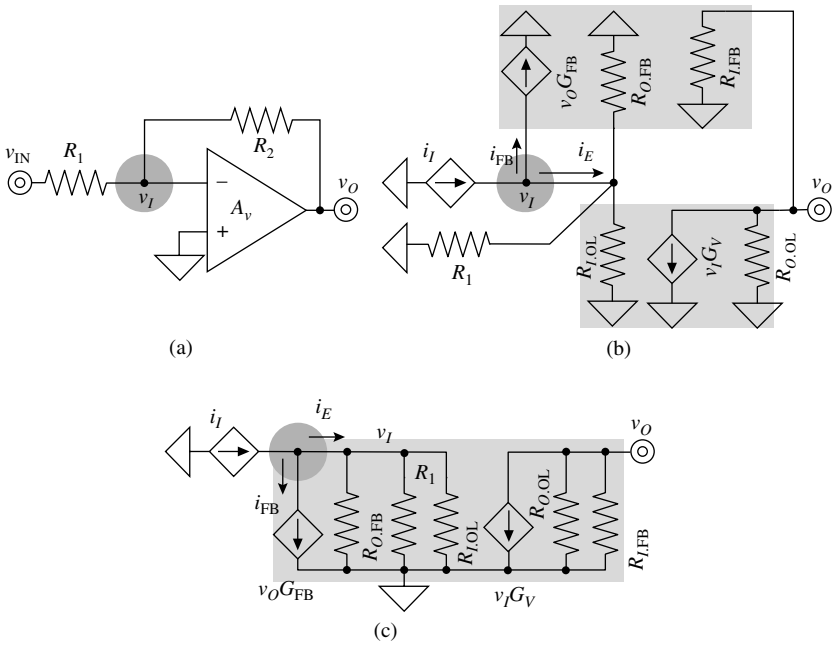


Figure 4.6 Shunt-mixing example: (*a*) the inverting op-amp configuration and (*b*) and (*c*) its equivalent two-port circuit decompositions.

equivalent circuit models for input resistor R_1, the op amp, and the feedback network, as shown in Fig. 4.6*b*, is beneficial because arbitrarily opening the loop may otherwise cloak the effects of impedances. In the case shown, the resistance looking back into R_1 is simply R_1 because input voltage source v_{IN} is assumed ideal, so v_{IN} and R_1 present Norton-equivalent source i_I and resistance R_1. Similarly, the op amp decomposes into Norton-equivalent transconductor source G_V, open-loop output resistance $R_{O.OL}$, and input resistance $R_{I.OL}$. In modeling the feedback network (i.e., R_2), it helps to combine the two-port equivalent circuits into one equivalent two-port model, as illustrated in Fig. 4.6*c*. In this combined circuit, the output resistance of the feedback network (i.e., $R_{O.FB}$) is a short-circuit parameter whose resistance is derived when v_O is zero (as not doing so would incorrectly expose $R_{O.FB}$ to the effects already modeled in G_{FB}), so $R_{O.FB}$ reduces to R_2:

$$R_{O.FB}\big|_{v_O = 0} \equiv \frac{v_I}{i_{R_2}} = R_2 \tag{4.25}$$

where i_{R_2} is the current flowing through R_2. Similarly, the feedback factor's transconductance G_{FB} is also a short-circuit parameter whose value is derived when v_i is zero and reduces to $1/R_2$:

$$G_{\text{FB}}\big|_{v_I = 0} \equiv \frac{i_{\text{FB}}}{v_O} = -\frac{1}{R_2} \tag{4.26}$$

Finally, and similarly, the input resistance of the feedback network $R_{I.\text{FB}}$ is again a short-circuit parameter whose resistance is derived when v_i is zero and reduces to R_2:

$$R_{I.\text{FB}}\big|_{v_I = 0} \equiv \frac{v_O}{i_{R_2}} = R_2 \tag{4.27}$$

The reason why no other impedance appears in $R_{O.\text{FB}}$, G_{FB}, and $R_{I.\text{FB}}$ is because other parameters in the two-port model already model them, which is why short- and open-circuit conditions in two-port models are imposed in the first place, to ensure no single parameter models the effects already captured by another.

Referring back to Fig. 4.6*b*, the forward open-loop transimpedance gain $A_{R.\text{OL}}$ is v_O/i_E, which, because $R_{I.\text{OL}}$ is normally substantially high and $R_{O.\text{OL}}$ considerably low, approximates to $-(R_1 \,||\, R_2)G_V R_{O.\text{OL}}$:

$$\begin{aligned} A_{R.\text{OL}} &\equiv \frac{v_O}{i_E} = -(R_1 \,||\, R_{I.\text{OL}} \,||\, R_{O.\text{FB}})G_V(R_{O.\text{OL}} \,||\, R_{I.\text{FB}}) \\ &= -(R_1 \,||\, R_{I.\text{OL}} \,||\, R_2)G_V(R_{O.\text{OL}} \,||\, R_2) \approx -(R_1 \,||\, R_2)G_V R_{O.\text{OL}} \end{aligned} \tag{4.28}$$

The feedback factor β_{FB} is i_{FB}/v_O or G_{FB}, which is $-1/R_2$. The closed-loop transimpedance gain $A_{R.\text{CL}}$ therefore reduces to roughly $-R_2$ (because $v_O = i_{\text{FB}}/\beta_{\text{FB}} \approx i_I/\beta_{\text{FB}}$):

$$A_{R.\text{CL}} = \frac{A_{R.\text{OL}}}{1 + A_{R.\text{OL}}\beta_{\text{FB}}} \approx \frac{-(R_1 \,||\, R_2)G_V R_{O.\text{OL}}}{1 + \left\{\dfrac{(R_1 \,||\, R_2)G_V R_{O.\text{OL}}}{R_2}\right\}} \approx -R_2 \tag{4.29}$$

The closed-loop input resistance is the loop-gain-reduced version of the effective input resistance of the circuit, which approximates to the loop-gain-reduced version of R_2:

$$R_{I.\text{CL}} = \frac{R_{I.\text{OL}} \,||\, R_1 \,||\, R_{O.\text{FB}}}{1 + A_{R.\text{OL}}\beta_{\text{FB}}} \approx \frac{R_1 \,||\, R_2}{1 + \left\{\dfrac{(R_1 \,||\, R_2)G_V R_{O.\text{OL}}}{R_2}\right\}} \approx \frac{R_2}{G_V R_{O.\text{OL}}} \tag{4.30}$$

which can be substantially low. Ultimately, the overall voltage gain of the circuit (i.e., A_V or v_O/v_{IN}) is the product of the translation of input voltage v_{IN} into i_I, which is the ohmic current flowing through the

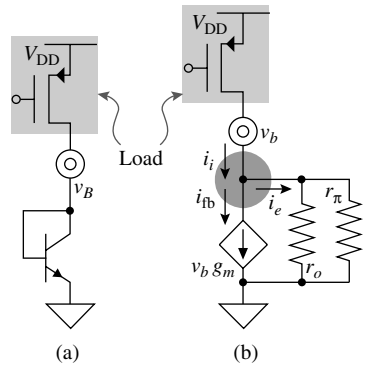


Figure 4.7 Shunt-mixing example: (*a*) a diode-connected transistor and (*b*) its small-signal equivalent circuit.

series combination of R_1 and $R_{I.CL}$, and closed-loop transimpedance gain $A_{R.CL}$ (i.e., v_O/i_I):

$$A_V \equiv \frac{v_O}{v_{IN}} = \left(\frac{i_I}{v_{IN}}\right)\left(\frac{v_O}{i_I}\right) \approx \left(\frac{1}{R_1 + R_{I.CL}}\right)(-R_{I.CL}) \approx -\frac{R_2}{R_1} \qquad (4.31)$$

assuming $R_{I.CL}$ is significantly lower than R_1. As expected, the result mimics that produced by applying the conventional op-amp approximations when connected in a negative-feedback configuration.

Though not often seen this way, the diode-connected transistor shown in Fig. 4.7 is also a shunt- or current-mixed negative-feedback circuit. A negative-feedback loop exists because, as base or gate voltage v_B or v_G increases, the transistor responds by shunting more current, whose net effect is to oppose the initial change in v_B or v_G. The input terminal leads into the base or gate, which may be a series mixer, except its emitter or source is at ground so the transistor's transconductor is not processing a differential small-signal voltage. In further identifying the mixer employed, it helps to decompose the circuit into its low-frequency small-signal equivalent (Fig. 4.7*b*), where the star current-shunting connection at the collector or drain is more apparent. Because there is only one node to this circuit and the input is already the current coming into the circuit, the output is base or gate voltage v_B or v_G. As such, the forward open-loop transimpedance gain $A_{R.OL}$ of the circuit is v_b/i_e or the parallel combination of r_π and r_o (i.e., $r_\pi \| r_o$) and its feedback factor β_{FB} is i_{fb}/v_b or transconductance g_m. Closed-loop transimpedance gain $A_{R.CL}$ therefore approximates to $1/g_m$,

$$A_{R.CL} = \frac{A_{R.OL}}{1 + A_{R.OL}\beta_{FB}} = \frac{r_\pi \| r_o}{1 + g_m(r_\pi \| r_o)} \approx \frac{1}{g_m} \qquad (4.32)$$

and closed-loop input resistance $R_{I.CL}$ to the loop-gain-reduced version of the effective open-loop input resistance of the circuit, which reduces to roughly $1/g_m$:

$$R_{I.CL} = \frac{R_{O.FB} \parallel R_{I.OL}}{1 + A_{R.OL}\beta_{FB}} = \frac{r_\pi \parallel r_o}{1 + g_m(r_\pi \parallel r_o)} \approx \frac{1}{g_m} \tag{4.33}$$

Again, as expected, the resulting closed-loop input resistance mimics the one derived in Chap. 3 for the diode-connected transistor in the basic mirror configuration.

Figure 4.8*a* illustrates a transistor-level embodiment of the more general shunt-mixed op-amp circuit in negative feedback. As with the series-mixed example, recognizing the transistor is a small-signal op-amp equivalent in shunt negative-feedback configuration and applying virtual-short/mirror approximations dictate small-signal base voltage v_b is equivalent to ac ground (i.e., $v_b \approx v_e = 0$) and input current i_{R1} equals feedback current i_{R2}. As such, the effective input current of the circuit is v_i/R_1, which is also equal to the small-signal current flowing through R_2, yielding an output voltage that is roughly $-v_iR_2/R_1$, that is,

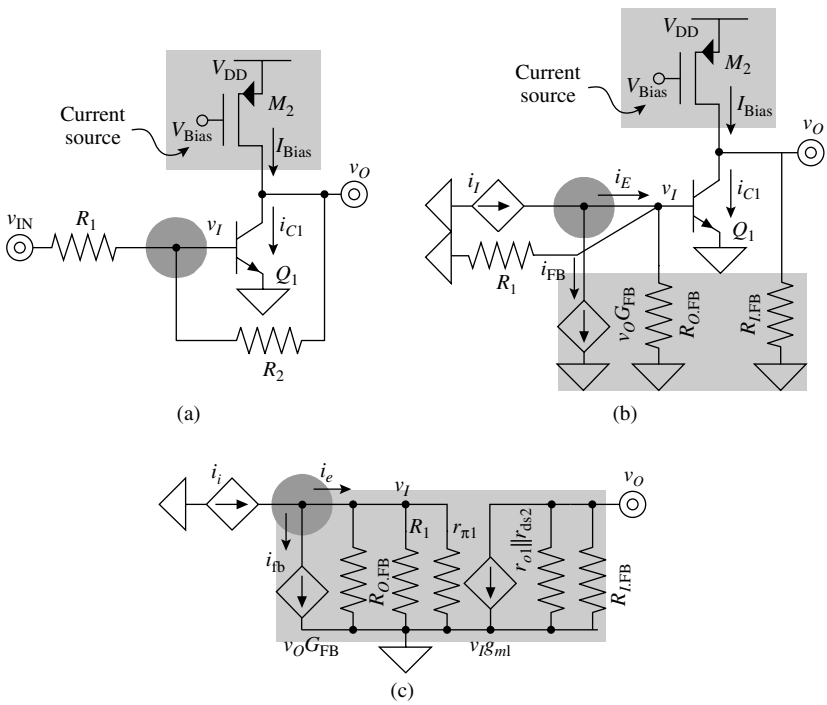


Figure 4.8 (*a*) Transistor-level embodiment of the shunt-mixed inverting op-amp case and (*b*) and (*c*) its equivalent two-port model decompositions.

a closed-loop gain of $-R_2/R_1$, the transistor-level equivalent of the inverting op-amp configuration gain.

In applying negative feedback theory, as before, it helps to decompose the circuit into its two-port equivalent models and further combine into a single collective two-port circuit to determine the feedback network's parameters, as shown in Fig. 4.8*b* and *c*. Not surprisingly, given the similarities between this circuit and the inverting op-amp configuration, the equivalent two-port model for the entire circuit resembles that of the op-amp counterpart, which means the feedback factor's input resistance $R_{I.FB}$, output resistance $R_{O.FB}$, and transconductance G_{FB} are R_2, R_2, and $-1/R_2$, respectively. Referring to Fig. 4.8*b*, the forward open-loop transimpedance gain $A_{R.OL}$, assuming the output is a voltage, is v_o/i_e or $-(r_{\pi 1} \,||\, R_1 \,||\, R_{O.FB})g_{m1}(r_{o1} \,||\, r_{ds2} \,||\, R_{I.FB})$ and the feedback factor β_{FB} is i_{fb}/v_o or G_{FB}, so the closed-loop transimpedance gain $A_{R.CL}$ is roughly $-R_2$:

$$A_{R.CL} = \frac{A_{R.OL}}{1 + A_{R.OL}\beta_{FB}} = \frac{-(r_{\pi 1} \,||\, R_1 \,||\, R_{O.FB})g_{m1}(r_{o1} \,||\, r_{ds2} \,||\, R_{I.FB})}{\left[1 + \dfrac{(r_{\pi 1} \,||\, R_1 \,||\, R_{O.FB})g_{m1}(R_{o1} \,||\, r_{ds2} \,||\, R_{I.FB})}{R_2}\right]} \approx -R_2 \qquad (4.34)$$

The closed-loop input resistance is the loop-gain-reduced version of the effective open-loop input resistance of the circuit, the latter of which is the parallel combination of $r_{\pi 1}$, R_1, and $R_{O.FB}$,

$$R_{I.CL} = \frac{r_{\pi 1} \,||\, R_1 \,||\, R_{O.FB}}{1 + A_{R.OL}\beta_{FB}} = \frac{r_{\pi 1} \,||\, R_1 \,||\, R_2}{\left[1 + \dfrac{(r_{\pi 1} \,||\, R_1 \,||\, R_2)g_{m1}(r_{o1} \,||\, r_{ds2} \,||\, R_2)}{R_2}\right]}$$

$$\approx \frac{R_2}{g_{m1}(r_{o1} \,||\, r_{ds2} \,||\, R_2)} \qquad (4.35)$$

which can be low. The overall small-signal voltage gain of the circuit (i.e., v_o/v_{in}) is the product of the ohmic translation of input voltage v_{in} to input current i_i and closed-loop transimpedance gain $A_{R.CL}$ of the negative feedback circuit (i.e., v_o/i_i), the result of which is approximately $-R_2/R_1$, as predicted by virtual-short approximations:

$$A_V \equiv \frac{v_o}{v_{in}} = \left(\frac{i_i}{v_{in}}\right)\left(\frac{v_o}{i_i}\right) \approx \left(\frac{1}{R_1 + R_{I.CL}}\right)(-R_2) \approx -\frac{R_2}{R_1} \qquad (4.36)$$

4.3 Samplers (Sensors)

Sampling a signal in the feedback path is the last basic component of a negative feedback loop. In sensing, the purpose of the sampler (shown in Fig. 4.1) is to measure a voltage or current with what amounts to a voltmeter or ammeter circuit, which is why *shunt* sampling refers to sensing voltages (because voltmeters are in *parallel*) and *series* sampling refers to sensing currents (because ammeters are in *series*), as illustrated in Fig. 4.9. Feedback factor β_{FB} feeds back sensed signal s_O to the mixer, forcing feedback signal s_{FB} to be a function of the sensed output. As with the mixer, series sampling, because the sensing circuit is in series, tends to increase the output resistance of an open-loop circuit and vice versa for shunt sampling.

4.3.1 Shunt (Voltage) Samplers

The sampling discussion starts with shunt sensing because voltages are perhaps easier to visualize than currents. Fundamentally speaking, a series voltage mixer is also a good voltage sampler, if the unused mixer terminal is not already in the feedback path, that is, if the output *is not mixed* into the loop (with a loop signal). As such, the same basic voltage mixers identified earlier in the series-mixing section (e.g., op amps, differential pairs, transistors, and difference transconductors) also apply to the foregoing shunt-sampling discussion, except output voltage v_O is at one of the differential inputs and the other is not in the feedback path, as shown in Fig. 4.10. Said differently, because the collector (or drain) current of a transistor is its driving output and its magnitude depends heavily on its base-emitter (or gate-source) voltage, bases (or gates) and emitters (or sources) are good voltage samplers, but only when the complementary terminal is either at ac ground or

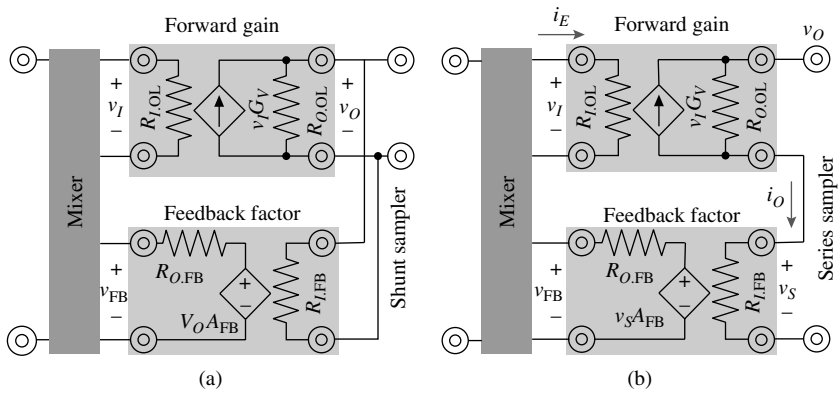
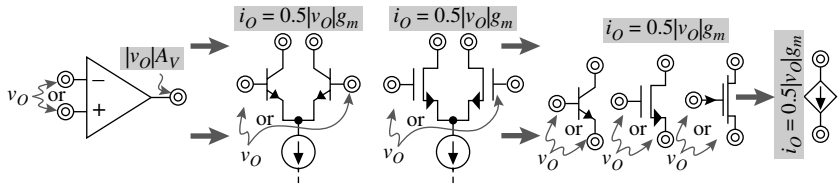


Figure 4.9 (*a*) Shunt and (*b*) series sampling in negative-feedback circuits.



*Unused terminal at 0, V_{Bias}, or v_I (but not in the feedback path)

Figure 4.10 Descending abstraction levels of most common shunt (voltage) samplers: op amp, differential pairs, transistors, and difference transconductors.

attached to v_I. Parenthetically, passive voltage dividers also present a good means of sensing a voltage and channeling its derivative to one of the above-mentioned sensors or directly into a mixer, so output signal v_O in Fig. 4.10 or feedback signal s_{FB} in Fig. 4.1 may also be voltage-divider fractions of the actual output voltage.

Just as shunt mixing decreases the open-loop input resistance of a circuit, shunt sampling reduces the open-loop output resistance by a loop-gain factor, where the open-loop resistance is the parallel combination of the feedback factor's input resistance $R_{I.FB}$ and the forward gain's output resistance $R_{O.OL}$. As a result, the closed-loop output resistance is the parallel combination of $R_{I.FB}$, $R_{O.OL}$, and the resistance introduced by two-port transconductance parameter G_V, which by recognizing that the forward open-loop gain A_{OL} or v_O/s_E is $G_V(R_{O.OL} \,||\, R_{I.FB})$, reduces to a loop-gain translation of $R_{O.OL} \,||\, R_{I.FB}$,

$$R_{G.CL} \equiv \frac{v_T}{i_T} = \frac{v_o}{i_G} = \frac{v_o}{v_o \beta_{FB} G_V} = \frac{R_{O.OL} \,||\, R_{I.FB}}{\beta_{FB} A_{OL}} \tag{4.37}$$

so

$$\begin{aligned} R_{O.CL} &= (R_{O.OL} \,||\, R_{I.FB}) \,||\, R_{G.CL} \\ &= (R_{O.OL} \,||\, R_{I.FB}) \,||\, \left(\frac{R_{O.OL} \,||\, R_{I.FB}}{\beta_{FB} A_{OL}} \right) = \frac{R_{O.OL} \,||\, R_{I.FB}}{1 + \beta_{FB} A_{OL}} \end{aligned} \tag{4.38}$$

In general, any time a shunting effect in a negative feedback exists, be it in the mixer or sampler, its effect on resistance is to reduce it by a factor of $1 + A_{OL}\beta_{FB}$. Since infinitesimally small output resistances constitute *ideal voltage sources* (because no loading resistance affects its output), many voltage supplies, references, and in some cases, op-amp output stages employ shunt-sampled negative feedback loops.

Perhaps the most easily recognized shunt samplers are op amps because their inputs can only process voltages. Two well-known examples are the noninverting and inverting op-amp configurations

shown in Figs. 4.4*a* and 4.6. In both cases, a resistor-divider network channels the output voltage to the inverting input of the op amp while the other input is outside the negative feedback loop with either incoming signal v_I, zero, or a bias voltage (i.e., ac ground). The output resistances of these op amps (in inverting and noninverting configurations) are the loop-gain translations of their open-loop counterparts:

$$R_{O.CL} = \left.\frac{R_{O.OL} \parallel R_{I.FB}}{1 + A_{V.OL}\beta_{FB}}\right|_{NI} \approx \frac{R_{O.OL} \parallel R_{I.FB}}{\left[1 + \dfrac{G_V(R_{O.OL} \parallel R_{I.FB})R_1}{R_2 + R_1}\right]} \approx \frac{R_2 + R_1}{G_V R_1} \tag{4.39}$$

for the noninverting series-mixed shunt-sampled feedback case and

$$R_{O.CL} = \left.\frac{R_{O.OL} \parallel R_{I.FB}}{1 + A_{R.OL}G_{FB}}\right|_{I} = \frac{R_{O.OL} \parallel R_{I.FB}}{1 + \dfrac{(R_{I.OL} \parallel R_1 \parallel R_{O.FB})G_V(R_{O.OL} \parallel R_{I.FB})}{R_{I.FB}}}$$

$$\approx \frac{R_{I.FB}}{(R_{I.OL} \parallel R_1 \parallel R_{O.FB})G_V} \approx \frac{R_2}{(R_1 \parallel R_2)G_V} = \frac{R_1 + R_2}{R_1 G_V} \tag{4.40}$$

for the inverting shunt-mixed shunt-sampled feedback counterpart, where subscript "NI" stands for noninverting and "I" for inverting, $A_{V.OL}$ is equivalent to $G_V(R_{O.OL} \parallel R_{I.FB})$, resistance $R_{I.OL}$ is assumed considerably large, and all relevant two-port parameters used (e.g., G_{FB}, $R_{I.FB}$, and $R_{O.FB}$) were previously derived in the mixer section.

Similarly, the general transistor-level circuits illustrated in Figs. 4.5*a* and 4.8, because they are transistor-level equivalents of the above-mentioned op amps, yield similar results, except their respective forward open-loop transconductances G_V differ. For the noninverting series-shunt instance, $R_{O.CL}$ is

$$R_{O.CL} = \left.\frac{R_{O.OL} \parallel R_{O.FB}}{1 + A_{O.OL}\beta_{FB}}\right|_{NI}$$

$$\approx \frac{R_{O.OL} \parallel R_{I.FB}}{\left[1 + \dfrac{(g_{m1}r_{ds3}g_{m2})(R_{O.OL} \parallel R_{I.FB})\left(R_1 \parallel \dfrac{2}{g_{m1}}\right)}{R_2 + \left(R_1 \parallel \dfrac{2}{g_{m1}}\right)}\right]}$$

$$\approx \frac{R_2 + \left(R_1 \parallel \dfrac{2}{g_{m1}}\right)}{(g_{m1}r_{ds3}g_{m2})\left(R_1 \parallel \dfrac{2}{g_{m1}}\right)} \tag{4.41}$$

where "NI" again stands for noninverting, R_{E1} shunts R_1 (and R_{E1} is not present in the general op-amp case), G_V is $g_{m1}r_{ds3}g_{m2}$, and $A_{V.OL}$ is equivalent to $G_V(R_{O.OL} \| R_{I.FB})$. $R_{O.CL}$ in the inverting shunt-shunt counterpart is

$$R_{O.CL} = \frac{R_{O.OL} \| R_{I.FB}}{1 + A_{R.OL}G_{FB}}\bigg|_I = \frac{R_{O.OL} \| R_{I.FB}}{1 + \dfrac{(R_{I.OL} \| R_1 \| R_{O.FB})g_{m1}(R_{O.OL} \| R_{I.FB})}{R_{I.FB}}}$$

$$\approx \frac{R_{I.FB}}{(R_{I.OL} \| R_1 \| R_{O.FB})g_{m1}} \approx \frac{R_2}{(r_{\pi 1} \| R_1 \| R_2)g_{m1}} = \frac{(r_{\pi 1} \| R_1) + R_2}{(r_{\pi 1} \| R_1)g_{m1}} \tag{4.42}$$

where "I" again stands for inverting, $R_{I.OL}$ is $r_{\pi 1}$, G_V is g_{m1}, and all relevant two-port parameters used were previously derived in the mixer section.

The diode-connected transistor in Fig. 4.7 is also a shunt-sampled circuit because the output voltage is at the base or gate terminal and its emitter or source counterpart is not in the feedback path. The output resistance is therefore the loop-gain-reduced version of $r_\pi \| r_o$:

$$R_{O.CL} = \frac{R_{O.OL} \| R_{I.FB}}{1 + A_{R.OL}\beta_{FB}} = \frac{r_\pi \| r_o}{1 + (r_\pi \| r_o)g_m} \approx \frac{1}{g_m} \tag{4.43}$$

where the parameters derived in the mixer section were used. Note the output in this case is also the input so $R_{I.CL}$ and $R_{O.CL}$ actually describe the same effect, which explains why they equal one another.

4.3.2 Series (Current) Samplers

Current samplers must necessarily conduct current so op amps and base and gate terminals are poor samplers, because little to no current flows through them. Collectors, drains, emitters, and sources, on the other hand, are better suited to conduct current (Fig. 4.11), and among these, collectors and drains are easier to identify because, if they constitute the sampling network, ruling out voltage sampling is readily justifiable. Emitters and sources, on the other hand, can sample voltages and currents, and they do the latter only when their respective bases and gates are in the feedback path, that is, when the current sampler

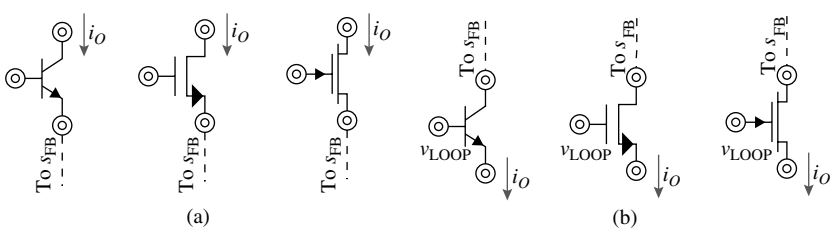


Figure 4.11 (*a*) Collector/drain and (*b*) emitter/source series (current) samplers.

also mixes the output voltage (in series) into the loop, which is complementary to the voltage-sampling conditions stated earlier in the shunt-sampler subsection. Note that feedback signal s_{FB} must ultimately be a function of the sensed output s_O (i.e., i_O) in the series-sampling case.

Because the sampler senses current, the negative feedback loop responds to oppose changes in output current as much as its loop gain allows, in effect regulating i_O against variations in output voltage. Infinitely high loop gains produce the characteristics of *ideal current sources*, which is why current regulators employ series samplers. The feedback loop, much like in the series-mixed case, increases the open-loop output resistance across the sensor by a loop-gain factor, the result of which, referring to Fig. 4.9, is the series combination of the feedback factor network's input resistance $R_{I.FB}$ and the loop-gain translation of the output resistance of the forward path $R_{O.OL}$. Assuming the input is a voltage, for simplicity, means feedback factor β_{FB} is v_{FB}/i_O (or $R_{I.FB}A_{FB}$) and forward gain A_{OL} is i_O/v_E (or G_V) and closed-loop output resistance $R_{O.CL}$ is therefore

$$R_{O.CL} \equiv \frac{V_T}{i_T} = \frac{v_O}{i_O} = \frac{v_{R_{O.OL}} + v_{R_{I.FB}}}{i_O} = \frac{(i_O + v_{R_{I.FB}} A_{FB} G_V) R_{O.OL} + v_{R_{I.FB}}}{i_O}$$

$$= \frac{[i_O + (i_O R_{I.FB}) A_{FB} G_V] R_{O.OL} + i_O R_{I.FB}}{i_O}$$

$$= (1 + \beta_{FB} A_{OL}) R_{O.OL} + R_{I.FB} \qquad (4.44)$$

as in the series-mixed case. Note series operations, in general, increase the resistance of a circuit.

One of the more classical series-sampling collector examples is the degenerated transistor discussed in Chap. 3 and illustrated in Figs. 4.4*b* and 4.12. The collector (or drain) is the output terminal, which is a poor

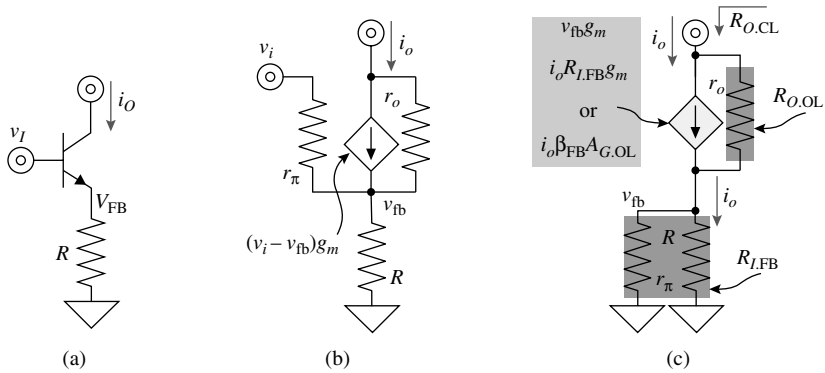


Figure 4.12 Emitter-degenerated common-emitter (CE) amplifier (*a*) circuit, (*b*) small-signal equivalent, and (*c*) short-circuit equivalent for deriving Norton-equivalent output resistance $R_{O.CL}$.

voltage sensor, the base (or gate) carries input voltage v_I, and the emitter (or source) is feedback voltage v_{FB}, all of which conforms to one of the conditions prescribed for a collector (or drain) current sensor in Fig. 4.11*a*. The overall Norton-equivalent output resistance of the circuit is a short-circuit parameter with v_I forced to zero, reducing its small-signal equivalent circuit (Fig. 4.12b) to the series combination of $r_\pi \,||\, R$ with the parallel combination of $v_{fb}g_m$ and r_o.

To determine the feedback parameters of the circuit, it helps to compare the reduced small-signal circuit to the output side of the general two-port series-sampled case in Fig. 4.9*b*. Doing so reveals that the parallel combination of r_π and R represents the feedback network's input resistance $R_{I.FB}$ and r_o the forward open-loop output resistance $R_{O.OL}$. The forward open-loop Norton-equivalent transconductance gain $A_{G.OL}$ or $i_o/(v_i - v_{fb})$ is also a short-circuit parameter where the voltage across r_o is zero, which means $A_{G.OL}$ is simply g_m. Feedback factor β_{FB} is v_{fb}/i_o or $r_\pi \,||\, R$. The resulting closed-loop output resistance $R_{O.CL}$ is the series combination of $R_{I.FB}$ or $r_\pi \,||\, R$ and the loop-gain increased version of $R_{O.OL}$ or r_o:

$$R_{O.CL} = (1 + A_{G.OL}\beta_{FB})R_{O.OL} + R_{I.FB} \approx [1 + g_m(R \,||\, r_\pi)]r_o + (R \,||\, r_\pi) \quad (4.45)$$

which resembles what was derived in Chap. 3 for the degenerated amplifier and the cascode current mirror, whose output is a degenerated common-emitter/source transistor. Because emitter current i_e exceeds collector current i_c (i.e., i_o) by the fraction flowing through the base, β_{FB} is larger than $r_\pi \,||\, R$, but not by much:

$$\beta_{FB} = \left(\frac{i_e}{i_o}\right)(r_\pi \,||\, R) = \left(\frac{\beta_{BJT} + 1}{\beta_{BJT}}\right)(r_\pi \,||\, R) \approx r_\pi \,||\, R \quad (4.46)$$

Figure 4.13 illustrates "amplified" extrapolations of the collector-sampling degenerated transistor case, as a voltage amplifier increases the loop gain and its resulting effect on the output resistance of the circuit. In the series-mixed version shown in Fig. 4.13*a*, for example, the op amp amplifies the forward open-loop transconductance gain $A_{G.OL}$ from g_m to $A_V g_m$, where A_V is the differential gain of the op amp, increasing the output resistance from roughly $r_{ds}(g_m R)$ to $r_{ds}(g_m A_V R)$. Figure 4.13*b* illustrates the shunt-mixed counterpart where input voltage v_{IN} is converted into a current before being mixed into the loop. Applying negative-feedback approximations dictate feedback current i_{FB} or i_O is a virtual mirror of input current i_I or $v_{IN}g_{m1}$. As a result, forward open-loop current gain $A_{I.OL}$ is i_o/i_e, or equivalently, M_C's $v_{gsC}g_{mC}$ divided by i_e:

$$A_{I.OL} = \frac{i_o}{i_e} = \frac{v_{gsC}g_{mC}}{i_e} = \frac{\{i_e r_{o1}[g_{mA}(r_{dsA} \,||\, r_{ds3})] - i_e r_{o1}\}g_{mC}}{i_e}$$

$$\approx r_{o1}g_{mA}(r_{dsA} \,||\, r_{ds3})g_{mC} \quad (4.47)$$

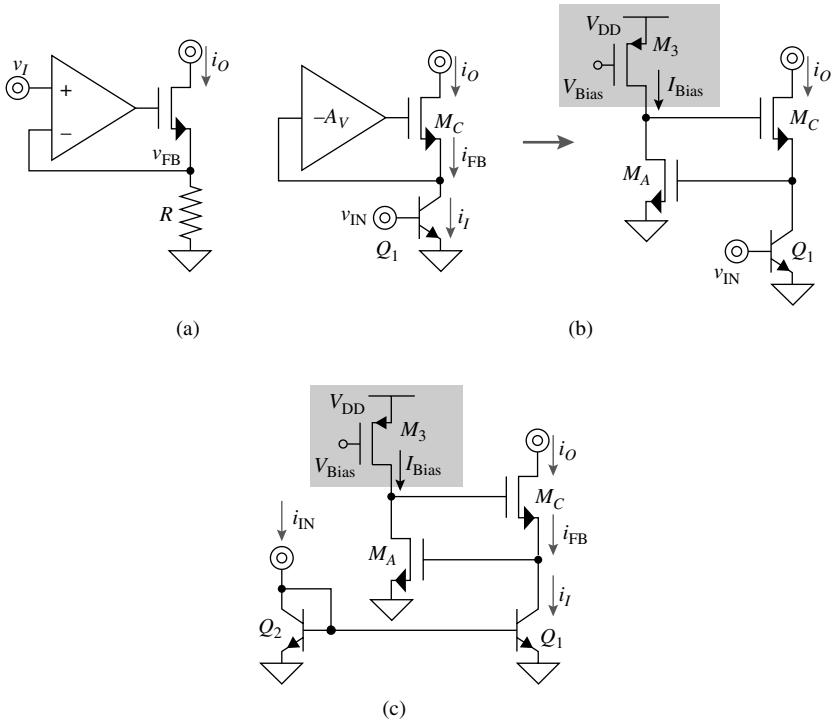


Figure 4.13 Amplified degenerated-transistor extrapolations: (*a*) amplified series-mixed series-sampled and (*b*) and (*c*) shunt-mixed series-sampled transformations, the latter of which is known as the regulated cascode current mirror.

and β_{FB} is i_{fb}/i_o or 1 A/A. The closed-loop output resistance of the circuit, which is evaluated when i_I is zero (as a two-port derivation), is approximately $r_{dsC}A_{I,OL}\beta_{FB}$ or $r_{dsC}r_{o1}A_V g_{mC}$, where A_V is $g_{mA}(r_{dsA} \,||\, r_{ds3})$.

The circuit in Fig. 4.13*c* represents a slightly modified version of the one in Fig. 4.13*b* where an NPN transistor that resides outside the loop (i.e., Q_2) preprocesses incoming current i_{IN} to define v_{IN} and ultimately set mixed input current i_I. The basic current mirror comprising Q_1 and Q_2 sets the value of mixed input current i_I to be approximately equal to i_{IN}. How i_I is mixed and i_O sampled in the loop, however, remain unaffected by the addition of Q_2 so the output resistance remains $r_{dsC}(g_{mC}A_V r_{o1})$. Note the circuit is a series combination of two negative feedback circuits: shunt-mixed shunt-sampled diode-connected transistor Q_2 and amplifier-enhanced shunt-mixed series-sampled degenerated transistor M_C. Because the amplification feature A_V allows the loop to better regulate i_O against output voltage variations and the basic function of the circuit is to mirror input current i_{IN}, the circuit is called a *regulated cascode current mirror*, a variation of the mirrors shown in Chap. 3 with higher output resistance.

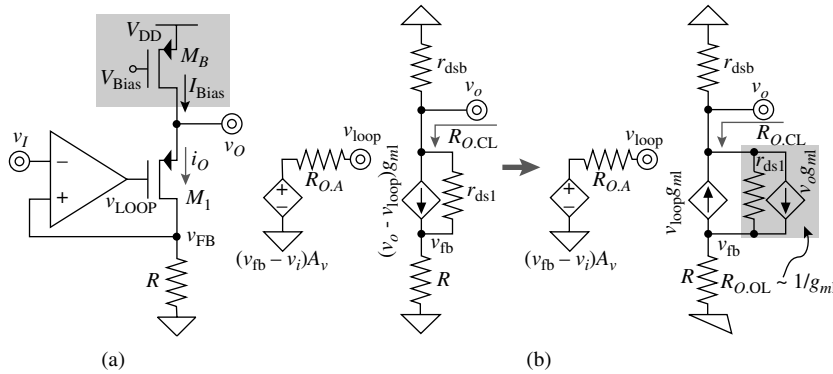


FIGURE 4.14 Series-mixed series-sampled source-sensing negative-feedback (*a*) circuit and (*b*) its small-signal equivalent circuit.

Figure 4.14 illustrates the use of a source (or emitter) sensor in a series-mixed series-sampled negative feedback circuit. The source of sensor transistor M_1 is a current (i.e., series) sampler because the source is the output terminal of the circuit and the corresponding gate is in the ac signal path—note that if output voltage v_{OUT} were an input, M_1 would series-mix v_{OUT} with v_{LOOP}. The output resistance of the entire loop (Fig. 4.14*b*) is the parallel combination of the resistance biasing transistor M_B presents (i.e., r_{dsB}) and the closed-loop output resistance of the negative-feedback circuit (i.e., $R_{O.CL}$); in other words, r_{dsB} is outside the negative feedback loop.

In comparing the output side of the general series-sampled negative-feedback circuit of Fig. 4.9*b* with this circuit (Fig. 4.14), the parallel combination of r_{ds1} and the equivalent resistance of transconductor g_{m1} reduces to $1/g_{m1}$ (from $(R+r_{ds1})/g_{m1}r_{ds1}$) and represents the forward open-loop output resistance $R_{O.OL}$, while R corresponds to the input resistance of the feedback network $R_{I.FB}$. Forward open-loop transconductance gain $A_{G.OL}$ is $i_o/(v_i - v_{fb})$ or $A_V g_{m1}/(1+g_{m1}r_{dsB})$ and, because all of output current i_o flows through the source, feedback factor v_{fb}/i_o or β_{FB} is R. The closed-loop output resistance $R_{O.CL}$ therefore approximates to $(1/g_{m1})[RA_V g_{m1}/(1+g_{m1}r_{dsB})]$ or roughly $RA_V/g_{m1}r_{dsB}$ and the overall output resistance of the circuit to $r_{dsB} \,||\, [RA_V/(g_{m1}r_{dsB})]$, which normally reduces to r_{dsB}. Note r_{ds1} and transconductor current i_{gm1}, as a degenerated common-source transistor, comprise a shunt-mixed negative-feedback loop, which is why $R_{O.OL}$ is low and approximately equal to $1/g_{m1}$.

If the base or gate of an emitter- or source-sampling transistor is not in the feedback path or mixed with the input, as shown in Fig. 4.15, unlike the series-sampling case of Fig. 4.14, the transistor senses a voltage because its transconductor current is a direct translation of its emitter or source voltage. Because dc biasing voltage V_{BIAS} is an ac ground in Fig. 4.15*a*, for example, small-signal transconductor current i_{gm} is

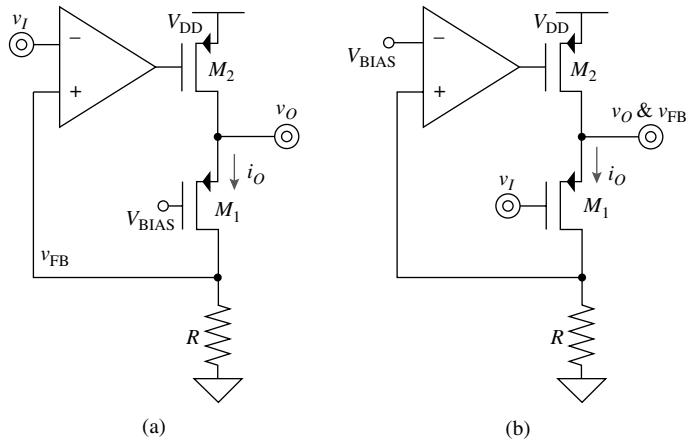


Figure 4.15 Series-mixed shunt-sampled source-sensor feedback circuits.

approximately $(v_o - 0)g_{m1}$ or $v_o g_{m1}$. Note the sampler, as in the voltage-mixing case, is the transconductor, not the entire transistor. Comparing the circuit to the output side of the shunt-sampled circuit of Fig. 4.9*a* shows that $R_{O.OL}$ is r_{ds2} and $R_{I.FB}$ the resistance into the source of M_1, which is $(r_{ds1}+R)/g_{m1}r_{ds1}$ or roughly $1/g_{m1}$. Similarly, on further inspection, $A_{V.OL}$ or $v_o/(v_i - v_{fb})$ is $(-A_V)(-g_{m2})(R_{O.OL} \,||\, R_{I.FB})$ or $A_V g_{m2}[r_{ds2} \,||\, (1/g_{m1})]$ or approximately $A_V g_{m2}/g_{m1}$ and β_{FB} or v_{fb}/v_o is $(1/R_{I.FB})R$ or roughly $g_{m1}R$. The closed-loop output resistance of the negative-feedback circuit is therefore approximately the loop-gain-reduced version of $1/g_{m1}$ or $1/(A_V g_{m2}R)g_{m1}$.

Injecting the input signal into the sensing transistor's base or gate, as illustrated in Fig. 4.15*b*, has similar effects on the output resistance because, as before, the transistor senses output voltage v_O. Since the mixer in the loop creates a virtual short between v_i and v_o (i.e., v_o is also feedback signal v_{fb}), loading the output has little effects on v_o, which is the same as saying the output resistance is low. Because v_i is zero when deriving $R_{O.CL}$ (as short-circuit two-port parameter derivations dictate), $R_{I.FB}$ is, as before, roughly $1/g_{m1}$ and $R_{O.OL}$ is r_{ds2}. In this case, forward open-loop gain $A_{V.OL}$ or $v_o/(v_i - v_{fb})$, because v_o is $(v_{fb} - v_i)g_{m1}RA_V[-g_{m2}(R_{O.OL} \,||\, R_{I.FB})]$, is $g_{m1}RA_V g_{m2}[r_{ds2} \,||\, (1/g_{m1})]$ or roughly $RA_V g_{m2}$ and feedback factor β_{FB} or v_o/v_{fb} is 1 V/V. The closed-loop output resistance of the circuit is therefore the loop-gain-reduced version of $R_{I.FB}$, that is, $R_{O.CL}$ is $1/g_{m1}RA_V g_{m2}$, as with the circuit in Fig. 4.15*a*.

4.4 Application of Negative-Feedback Theory

The power behind negative feedback resides on its resulting approximations and loop-gain effects. Memorizing circuit topologies and equations help but designing beyond the state of the art demands

deeper understanding. The circuits presented in the previous sections aim to exemplify the fundamentals of negative feedback with the three-terminal transistor as its core, as all circuits ultimately decompose into transistors and their constituent small-signal parameters. The most basic, yet most powerful feature of a negative-feedback circuit is it regulates (i.e., controls) its output (when its loop gain is considerably larger than 1) so that the difference between its mixing inputs is negligibly small, virtually short-circuiting or mirroring the input signals. The closed-loop gain of the circuit at any point is therefore, and simply, the gain translation from the virtual short/mirror point. For instance, output s_O is approximately the ratio of s_{FB} (or s_I) and feedback factor β_{FB} (i.e., $s_O = s_{FB}/\beta_{FB} \approx s_I/\beta_{FB}$ or $A_{CL} \approx 1/\beta_{FB}$) and, similarly, error s_E the ratio of s_{FB} and loop gain $A_{OL}\beta_{FB}$ (i.e., $s_E = s_{FB}/A_{OL}\beta_{FB} \approx s_I/A_{OL}\beta_{FB}$).

Shunt feedback, whether it is at the input or output, decreases the impedance of the circuit. At the input, for example, the loop responds to ensure a virtual mirror exists between input and feedback currents i_I and i_{FB}, producing a substantially small error current i_E and therefore a negligibly small input voltage v_I (across $R_{I.OL}$ and $R_{O.FB}$). Irrespective of the magnitude of i_I, v_I is small, which is equivalent to saying the negative-feedback circuit offers considerably low input impedance. In the case of shunt sampling, the loop ensures sensed output voltage v_O remains unchanged in the face of changing load currents. In other words, shunt sampling regulates v_O against variations in output current i_O, which has the same effects of a change in i_I on v_I in the shunt-mixing example—little to no effect on v_O— so the circuit presents a substantially low output impedance.

Conversely, series mixing and sampling increase the impedance of a circuit. In the case of series mixing, for instance, the loop responds to ensure feedback voltage v_{FB} equals input voltage v_I, producing a substantially small error voltage v_E whose implied ohmic current through input resistance $R_{I.OL}$ is also considerably small (noting i_I is $v_E/R_{I.OL}$). The small variation that results in input current i_I when presented with an input-voltage excursion is the manifestation of high input resistance. Similarly, sensing and therefore regulating output current i_O (as in series sampling) against variations in output voltage v_O results in small changes in i_O in spite of considerable changes in v_O, in other words, produces high output impedance.

In applying these generalized conclusions, however, it is important to identify the mixing and sampling functions. Fundamentally, voltage mixers (i.e., series) and samplers (i.e., shunt) reduce to base-emitter and gate-source terminals. Voltage mixers result when input voltage v_I is at one terminal and a loop signal (i.e., v_{FB} or v_{LOOP}) is at the other and voltage samplers when output voltage v_O is at one terminal and the other is anything but a loop signal (i.e., v_I or ac ground). Current (i.e., shunt) mixers result from star-mixing connections between input current i_I, feedback current i_{FB}, and error current i_E into base/gate or emitter/source terminals. Because collectors/drains and emitters/sources carry the driving

current of a transistor, collectors/drains and emitter/sources are good current samplers, but the latter only when their respective bases/gates are in the feedback path with a loop voltage (i.e., v_{LOOP}) driving them.

Ultimately, determining the circuit's loop gain is important to quantify the effects of feedback on resistance. Determining forward open-loop gain A_{OL} and feedback factor β_{FB} independently, however, may not be necessary, not even for stability, which is entirely a function of loop gain $A_{OL}\beta_{FB}$. The only motivation for evaluating their values independently is to ascertain the closed-loop gain of the circuit, and with the simplifying statements already presented in this regard, not even then. In practice, when inspecting a circuit, it is often easier to derive loop gain $A_{OL}\beta_{FB}$ than its constituent A_{OL} and β_{FB} parameters because the latter involves decomposing the circuit into its two-port equivalent models. The absolute value of this loop gain is the gain across the loop when mixed input i_I or v_I in the mixing transconductor is zero, irrespective of where the starting point is relative to the input or output terminals of the circuit. A shunt resistance would then be the loop-gain-reduced version of its open-loop resistance or approximately $R_{OL}/(A_{OL}\beta_{FB})$ and the series resistance the loop-gain-increased version of its open-loop resistance or roughly $R_{OL}(A_{OL}\beta_{FB})$.

The easiest means of quantifying the loop gain is to start with a voltage at a high-resistance node like a gate, base, or op-amp input terminal and traverse through the loop back to that point. In doing so, as already mentioned, the mixed input must be zero. This is not always easy to do at the transistor level because the mixer is inside the transistor and zeroing a transistor's terminal may, and often does, camouflage the effects of transistor impedances r_o or r_{ds}, r_π, C_π or C_{GS}, and/or C_μ or C_{GD}. The best way to zero a voltage-mixed input v_I is to decompose the mixing transistor or op amp into its small-signal equivalent and zero the mixed-input component of the dependent source that is mixing s_I and s_{FB} (e.g., zero v_I in $(v_I - v_{FB})g_m$ or $(v_I - v_{FB})A_V$ to yield $-v_{FB}g_m$ or $-v_{FB}A_V$). In shunt-mixed cases, decomposing the incoming signal into its Norton-equivalent circuit and zeroing or discarding only the dependent current source, which is the one that carries current input i_I, is usually the most straightforward means of opening the loop. Note the mixer inverts the signal and the resulting gain across the loop is actually an inverted translation of the loop gain: $-A_{OL}\beta_{FB}$.

In practice, negative-feedback analysis produces, for the most part, reasonable approximations, not exact relationships. To start, the two-port equivalent models used to decipher negative-feedback effects only estimate and mimic the response of a circuit to first order. To make matters worse, more approximations result when applying feedback theory to intertwined loops, such as emitter- and source-degenerated transistors that also double as mixers or samplers. In such cases, it is prudent to assume the higher loop gain overwhelms the others, but only if the ratio between the two loop gains is sufficiently high, like greater than 10.

Below is a succinct and concise, though brief, summary of important conclusions drawn from the foregoing discussion:

1. Feedback signal s_{FB} is the virtual short/mirror of s_I: $s_{FB} \approx s_I$.
2. Output signal s_O (and A_{CL}) is the β_{FB} translation of the virtual short/mirror: $s_O = s_{FB}/\beta_{FB} \approx s_I/\beta_{FB}$.
3. Shunt feedback reduces the open-loop resistance by a factor of $1 + A_{OL}\beta_{FB}$.
4. Series feedback increases the open-loop resistance by a factor of $1 + A_{OL}\beta_{FB}$.
5. Transistor's transconductance g_m is *both* a good voltage mixer and sampler.
6. Current mixers are star connections of i_I, loop signal i_{FB}, and error current i_E.
7. Collector/drain and emitter/source terminals are good current samplers, but the latter only when their respective base/gate terminals carry loop signals.
8. In opening a loop to determine loop gain $A_{OL}\beta_{FB}$, zero s_I *only* in the mixing current/voltage source.
9. Feedback approximations include: (a) the use of two-port models, (b) the assumption that the effects of the higher loop gain of multiple intertwined loops overwhelm the others', and (c) small-signal impedance approximations.

4.5 Stability

A positive-feedback loop, instead of working against the conditions that force variations across the input terminals of its mixer, helps external forces increase their differentiating impact on what would have otherwise been a virtual short or mirror. This positive-feedback effect is mathematically apparent when loop gain $A_{OL}\beta_{FB}$ reaches unity-gain frequency f_{0dB} with a total shift in phase of 180° (i.e., $A_{OL}\beta_{FB}$ at f_{0dB} is −1, as shown in Fig. 4.16*a*). At this point and under these conditions, there is positive feedback at f_{0dB} and A_{CL} explodes to infinity at f_{0dB} (as illustrated in Fig. 4.16*b*) because A_{CL}'s denominator (i.e., $1 + A_{OL}\beta_{FB}$) approaches zero:

$$A_{CL} = \left.\frac{A_{OL}}{1 + A_{OL}\beta_{FB}}\right|_{A_{OL}\beta_{FB}=1\angle 180^o} = \frac{A_{OL}}{1-1} \to \infty \tag{4.48}$$

The criterion for a feedback circuit to remain stable is therefore to ensure its loop gain LG or $A_{OL}\beta_{FB}$ (not forward gain A_{OL} alone) has less than 180° of phase shift at the loop gain's f_{0dB}. Because of this unstable

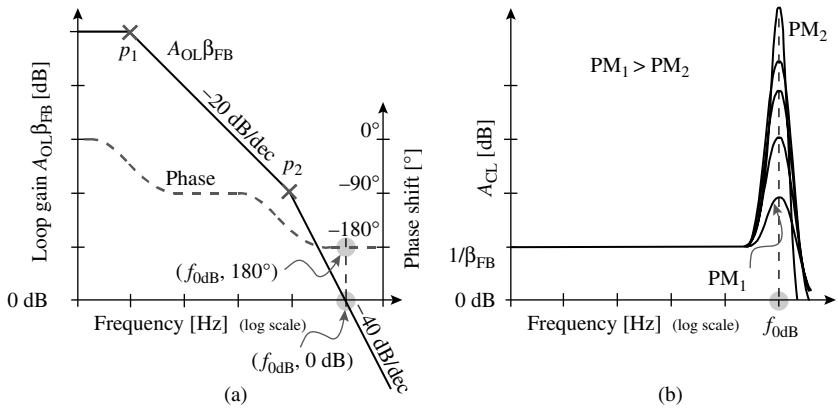


Figure 4.16 (*a*) Open-loop gain and phase (i.e., Bode plot) and (*b*) resulting closed-loop gain responses of a negative-feedback circuit as phase margin (PM) decreases to its unstable state of 0°.

point, *phase margin* (*PM*) refers to how much margin in phase exists at f_{0dB} before reaching 180°. Similarly, *gain margin* (*GM*) is how much margin there is in gain below the 0 dB axis before reaching the frequency where the loop gain incurs 180° of phase shift. As a result, the closed-loop circuit's proneness to instability (i.e., peaking effects) increases with decreasing phase and gain margins PM and GM, as illustrated in Fig. 4.16*b*.

Each pole shunts incoming signals to ac ground at a rate of 20 dB per decade (i.e., linearly) past the pole's location as frequency increases and phase-shifts (i.e., delays) signals by approximately −90° a decade past it, −45° at its location, and 0° a decade before it, as illustrated in Fig. 4.16*a* for poles p_1 and p_2. A zero, on the other hand, feeds forward a signal and increases its magnitude by 20 dB per decade past the zero's location and phase-shifts it by +90°, +45°, and 0° a decade past, at, and a decade before it. Similarly, a right-half-plane (RHP) zero also feeds forward a signal and increases it by 20 dB per decade, like a left-half-plane (LHP) zero, but like a pole, phase-shifts it by −90°, −45°, and 0° a decade past, at, and a decade before it.

Figure 4.16*a* graphically illustrates the Bode-plot response of a two-pole system whose pole locations precede f_{0dB} by at least one decade so PM is zero, which constitutes the makings of an unstable system. Because each pole phase-shifts a signal −90°, there is already 180° of phase shift by the time the loop gain crosses the 0 dB axis, resulting in zero phase and gain margins. Had the second pole (i.e., p_2) been within a decade of f_{0dB}, there would have been less than 180° of phase shift at f_{0dB} and more phase and gain margins, attenuating the closed-loop gain's peaking effect in Fig. 4.16*b*, which is a manifestation of instability or growing oscillations at peaking frequency f_{0dB}.

4.6 Frequency Compensation

The general strategy for stabilizing negative-feedback circuits is to ensure the loop gain (i.e., $A_{OL}\beta_{FB}$) approaches the unity-gain frequency (i.e., f_{0dB}) at a rate of −20 dB per decade, as illustrated in Fig. 4.17, while keeping all right-half-plane (RHP) zeros at considerably higher frequencies. Conventionally, a single dominant low-frequency pole p_1 is established, a secondary pole p_2 placed near or above f_{0dB}, and all remaining parasitic poles placed at least a decade above f_{0dB}, ensuring an overall phase margin of 45° or greater. Although less often the case, designers may allow one or two in-band poles and use left-half-plane (LHP) zeros to cancel their shunting effects within a decade below f_{0dB}. This latter technique is less popular because the system is more prone to instabilities during start-up conditions, when the gain and its respective poles and zeros shift before reaching their steady-state locations. Besides, a rather useful result of a single-pole response is that the *gain-bandwidth product* (*GBW*) along the −20 dB per decade drop is constant and equal to unity-gain frequency f_{0dB} because the gain decreases by the same factor the frequency increases (i.e., $A_{OL}\beta_{FB}$ is linearly proportional to $1/s$, which is proportional to $1/f$ past dominant pole p_1). Because regulators suffer from relatively extreme variations in load, however, allowing one pole and offsetting its effects with a LHP zero is more acceptable in regulators.

A circuit's response to a step input change is a time-domain manifestation of phase shift and phase margin PM. More explicitly, delay represents phase shift and overall settling time (as the output oscillates about its target until it converges) corresponds to PM, which means multiple poles further delay the circuit and lower PM values

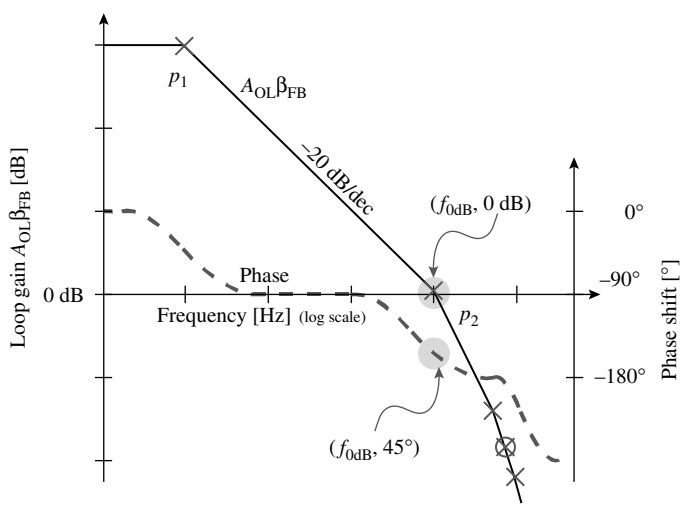


Figure 4.17 Stable Bode-plot response of a negative feedback circuit.

increase oscillations and settling time. On one extreme, zero PM prevents the output from settling to its intended target, and on the other, 90° of PM eliminates all oscillations in the response, allowing the output to approach its target without overshooting it. Similarly, 45° of PM produces less than three oscillating rings in the output before settling within 10% of its target, extending settling time beyond its 0–90% delay by approximately three f_{0dB}-equivalent periods (i.e., f_{0dB} is $1/T_{0dB}$). Although not always the case, a PM target of 45° is a popular design objective for feedback circuits.

4.6.1 Establish a Dominant Pole

Introduce a Low-Frequency Pole

Perhaps the easiest, though not always the optimum means of compensating a negative feedback circuit is to *add* a pole at frequencies well below the location of all other poles, as shown with additional pole p_A in Fig. 4.18*a*. There are two basic drawbacks to this technique the most important of which is reduced bandwidth, that is, a considerably lower compensated unity-gain frequency (i.e., $f_{0dB.A}$ is considerably below uncompensated f_{0dB}). Feedback loops with higher compensated unity-gain frequencies maintain desirable negative-feedback effects on gain and impedances through higher frequencies, thereby enabling them to process higher frequency signals, and compensation of any form, unfortunately, necessarily reduces the operating bandwidth from its uncompensated state. The extent to which compensation reduces f_{0dB} indicates the relative efficacy of the strategy against other approaches. The other drawback of adding a pole to the circuit may be power, as adding a pole usually requires more quiescent-current flow because the mere addition of a node with nonzero finite resistance, to say nothing of the circuitry that surrounds it, dissipates power when confronted with incoming signals.

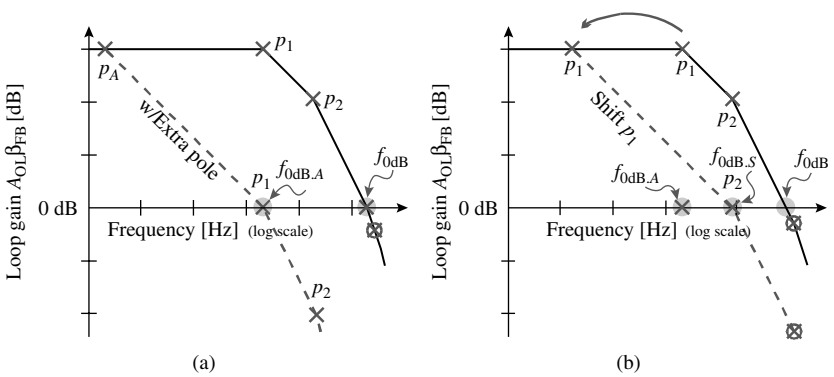


Figure 4.18 Bode plots for compensated negative-feedback circuits after (*a*) adding pole p_A to the response and (*b*) pulling lowest frequency pole p_1 to lower frequencies.

A better means of achieving reasonable phase-margin response is by shifting an already existing pole to lower frequencies, as demonstrated with p_1 in Fig. 4.18*b*. There are two basic advantages to this approach. First, pulling a pole to low frequencies is a relatively easy solution because doing so amounts to adding a shunt capacitor from a high-resistance node to ac ground. Second, there is one less pole to consider, altogether avoiding the drop in gain and shift in phase that extra pole p_A would have otherwise incurred. The net result is that the compensated shift-assisted unity-gain frequency $f_{0\text{dB}.S}$ can reach higher frequencies than its pole-added counterpart $f_{0\text{dB}.A}$ (Fig. 4.18*b*).

Miller Compensation

One of the most convenient methods of establishing a low-frequency pole is to split two poles with Miller compensation, pulling the lowest frequency pole to lower frequencies and pushing the next dominant pole to higher frequencies. From a feedback perspective, adding a Miller capacitor C_M across an inverting stage, as shown in Fig. 4.19*a*, establishes a negative feedback loop mimicking the inverting op-amp configuration of Fig. 4.6*a*, where C_M replaces feedback resistor R_2. The resulting i_I, i_{FB}, and i_E star connection shunt-mixes the input and therefore produces a loop-gain-reduced version of the open-loop input impedance that is approximately equal to $1/(sC_MA_V)$, or equivalently, an input capacitance C_{IN} that is roughly A_VC_M:

$$Z_{I.\text{CL}} = \frac{Z_{I.\text{OL}} \parallel Z_{O.\text{FB}}}{1 + A_{R.\text{OL}}\beta_{\text{FB}}} = \frac{Z_{I.\text{OL}} \parallel Z_{O.\text{FB}}}{1 + [(Z_{I.\text{OL}} \parallel Z_{O.\text{FB}})A_V]\left(\dfrac{1}{Z_{\text{FB}}}\right)}$$

$$= \frac{Z_{O.\text{FB}}}{1 + A_V} = \frac{1}{(1 + A_V)(sC_M)} \equiv \frac{1}{sC_{\text{IN}}} \tag{4.49}$$

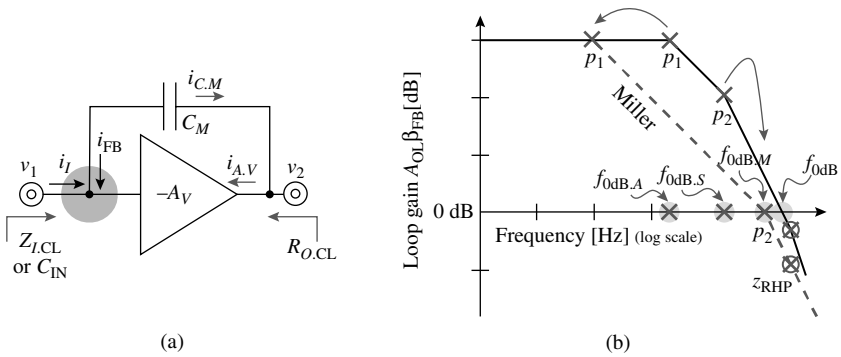


Figure 4.19 (*a*) Using Miller capacitor C_M to (*b*) split input and output poles p_1 and p_2 and compensate an otherwise unstable feedback circuit.

As in the inverting op amp of Fig. 4.6*a*, the output is shunt-sampled and the closed-loop output impedance $Z_{O.CL}$ at higher frequencies, when C_M is for all practical purposes a short circuit, is therefore the loop-gain-reduced version of the open-loop resistance, or the reciprocal of the op amp's forward open-loop transconductance gain G_{OL}:

$$Z_{O.CL} = \frac{R_{O.OL} \parallel Z_{I.FB}}{1 + A_{R.OL}\beta_{FB}} = \frac{R_{O.OL} \parallel Z_{I.FB}}{1 + [(Z_{I.OL} \parallel \mid Z_{I.FB})G_{OL}(R_{O.OL} \parallel Z_{I.FB})]\left(\frac{1}{Z_{FB}}\right)}$$

$$= \left.\frac{R_{O.OL} \parallel \left(\frac{1}{sC_M}\right)}{1 + \left(R_{I.OL} \parallel \frac{1}{sC_M}\right)G_{OL}\left[R_{O.OL} \parallel \left(\frac{1}{sC_M}\right)\right]sC_M}\right|_{f >> \frac{1}{2\pi R_{I.OL}C_M}}$$

$$\approx \frac{1}{G_{OL}} \tag{4.50}$$

at or above frequencies where impedance $1/sC_M$ is considerably lower than $R_{I.OL}$ (i.e., past $1/2\pi R_{I.OL}C_M$). Note the loop is open at low frequencies (that is, impedance $1/sC_M$ is considerably high) so the results apply at or above moderate frequencies, when C_M effectively shorts, which are the same results obtained in the Miller-effect discussion of the common-emitter/source amplifier in Chap. 3.

As a result, intentionally placing a Miller capacitor across an inverting gain stage in the negative-feedback path pulls its input pole p_1 to lower frequencies by presenting an amplified capacitance C_{IN} to v_1 and pushes its output pole p_2 to higher frequencies by presenting a low-impedance point $Z_{O.CL}$ to v_2. Pole p_1 can therefore become the dominant, low-frequency pole of the loop, as shown in Fig. 4.19*b*, and p_2 the secondary pole. The resulting compensated unity-gain frequency $f_{0dB.M}$ is lower than its uncompensated state f_{0dB} but higher than in the previous two approaches: introducing low-frequency pole p_A (i.e., $f_{0dB.M} > f_{0dB.A}$) and only shifting p_1 to lower frequencies (i.e., $f_{0dB.M} > f_{0dB.S}$). The drawback of Miller capacitor C_M is that it also presents an out-of-phase feed-forward path to the output. This is problematic when feed-forward capacitor current $i_{C.M}$ is equal to or exceeds the amplifier's inverting current (i.e., $i_{A.V}$) because it inverts the polarity of the signal. This inversion occurs at relatively high frequencies, above RHP zero z_M, which is approximately at $G_{OL}/2\pi C_M$:

$$i_{C.M} = (v_I - v_O)sC_M = (v_I - 0)sC_M\Big|_{z_M = \frac{G_{OL}}{2\pi C_M}} \equiv i_{A.V} = v_I G_{OL} \tag{4.51}$$

where v_o is zero because, when $i_{C.M}$ equals $i_{A.V}$, no current flows through the op amp's Norton-equivalent output resistance $R_{O.OL}$. In any case, Miller-compensated unity-gain frequency $f_{0dB.M}$ must remain below z_{RHP}.

4.6.2 Right-Half-Plane (RHP) Miller Zeros

Shift the RHP Zero

Splitting the first two dominant poles in the feedback loop with a Miller capacitor is, of course, ideal in compensating a negative feedback circuit, but the trailing RHP zero limits how high the compensated unity-gain frequency (i.e., $f_{0dB.M}$) can reach. This limitation is typically more acute in MOS circuits because the transconductances of square-law devices fall short of their exponential bipolar counterparts. A resistor in series with Miller capacitor C_M, however, as shown in Fig. 4.20*a*, increases the feed-forward impedance through C_M and therefore *limits* (i.e., impedes) feed-forward current $i_{C.M}$, shifting RHP Miller zero z_M to higher frequencies,

$$i_{C.M} = \frac{(v_{in} - v_{out})}{\frac{1}{sC_M} + R_M} = \left.\frac{(v_{in} - 0)}{\frac{1}{sC_M} + R_M}\right|_{z_M = \frac{1}{2\pi C_M\left|\frac{1}{G_{OL}} - R_M\right|}} \equiv i_{A.V} = v_{in} G_{OL} \qquad (4.52)$$

Equating *nulling* Miller resistor R_M to $1/G_{OL}$ shifts z_M to infinitely high frequencies, except equating these two parameters exactly across process and temperature corners is not easy in practice. However, a larger R_M (i.e., greater than $1/G_{OL}$) shifts the zero into the left-half plane (LHP), where it could potentially help increase phase margin.

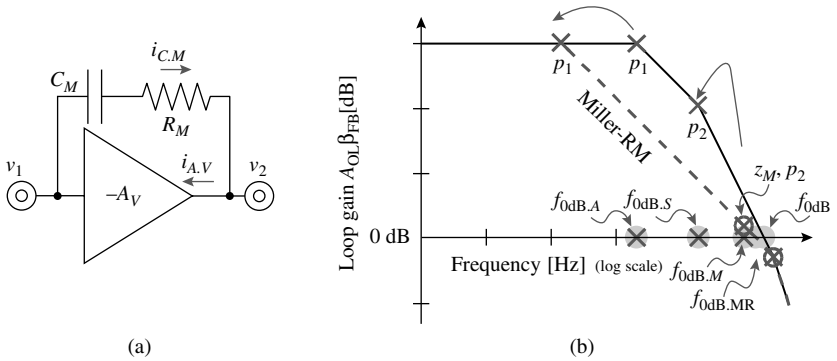


Figure 4.20 (*a*) Using "nulling" Miller resistor R_M to limit (i.e., impede) out-of-phase feed-forward current $i_{C.M}$ and (*b*) pull Miller zero z_M to the left half of the *s* plane to offset the signal-shunting effects of secondary pole p_2.

In fact, this LHP zero can cancel the effects of secondary pole p_2 (Fig. 4.20*b*) and allow compensated resistor-assisted Miller unity-gain frequency $f_{0dB.MR}$ approach its uncompensated counterpart of f_{0dB}. Again, matching z_M and p_2 is not trivial, forcing the designer to decrease $f_{0dB.MR}$ slightly below f_{0dB} for margin. Pulling z_M below p_2 is risky because z_M could extend $f_{0dB.MR}$ into the parasitic-pole region, where phase drops quickly and unstable conditions are more likely to occur.

Eliminate the RHP Zero

Fundamentally, splitting the primary and secondary poles with a Miller capacitor is strictly a feedback effect, as shunt mixing increases the effective input capacitance of the circuit and shunt sampling decreases the output impedance. Consequently, ensuring the ac-signal path through the capacitor is unidirectional in the feedback sense retains the feedback pole-splitting features of C_M while eliminating the feed-forward path and the right-half-plane zero that results (i.e., z_M). Driving C_M with unidirectional unity-gain voltage buffers in the feedback sense, as shown in the general case and the transistor-level common-drain (CD) follower embodiment of Fig. 4.21*a* and *b*, impress v_O on the input terminal of C_M while *blocking* feed-forward current into v_O. Similarly, channeling capacitor current $i_{C.M}$ through a unidirectional current buffer, as illustrated with the common-gate (CG) current buffer of Fig. 4.21*c*, steers capacitor feedback current to the input mixer while not allowing feed-forward current to flow back through the buffer to v_O because the output resistance of the buffer, as seen from v_I, is substantially high (i.e., the output drain resistance of a source-degenerated transistor).

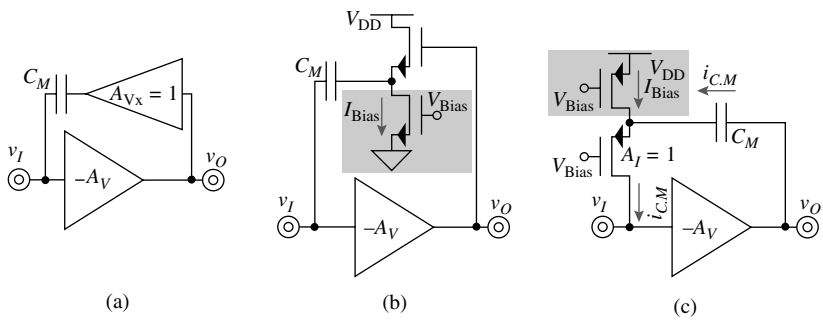


Figure 4.21 Eliminating the feed-forward right-half-plane zero (i.e., z_M) a Miller capacitor introduces with (*a*) a voltage-buffer and (*b*) a common-drain (CD) source-follower embodiment and (*c*) a common-gate (CG) current buffer.

4.6.3 Multiplying the Miller Effect

Replacing the unity-gain buffers in Fig. 4.21 with amplifiers further multiplies the Miller effects by increasing the loop gain of the circuit. In the case of the voltage buffer, however, increasing its gain in practice is more cumbersome because the high-impedance node of the additional gain stage in the feedback network (e.g., common-emitter/source amplifier) often presents a relatively low-frequency pole and controlling its biasing point is not always straightforward. Increasing the current gain, on the other hand, as illustrated in Fig. 4.22, is more viable because it tends to introduce higher frequency poles and offer more stable biasing conditions. The basic idea is to channel displacement capacitor current $i_{C.M}$ into a current amplifier and applying its output to v_I, as the transistor-equivalent shown in Fig. 4.22*b* does through a pair of amplifying current mirrors, where the second mirror negates the inverting effect of the first mirror. The resulting closed-loop input capacitance is a multiplied version of the Miller counterpart and approximately equal to $A_I A_V C_M$ (where A_I is the total current gain through the amplifying current mirrors) and the closed-loop output impedance at high frequencies is a reduced version of its Miller counterpart at roughly $1/A_I G_{OL}$:

$$Z_{I.CL} = \frac{Z_{I.OL} \parallel Z_{O.FB}}{1 + A_{R.OL}\beta_{FB}} = \frac{Z_{I.OL} \parallel Z_{O.FB}}{1 + [(Z_{I.OL} \parallel Z_{O.FB})A_V]\left[\left(\frac{1}{Z_{O.FB}}\right)A_I\right]}$$

$$= \frac{Z_{O.FB}}{1 + A_V A_I} \approx \frac{1}{A_V A_I}\left(\frac{1}{sC_M}\right) \equiv \frac{1}{sC_{IN}} \tag{4.53}$$

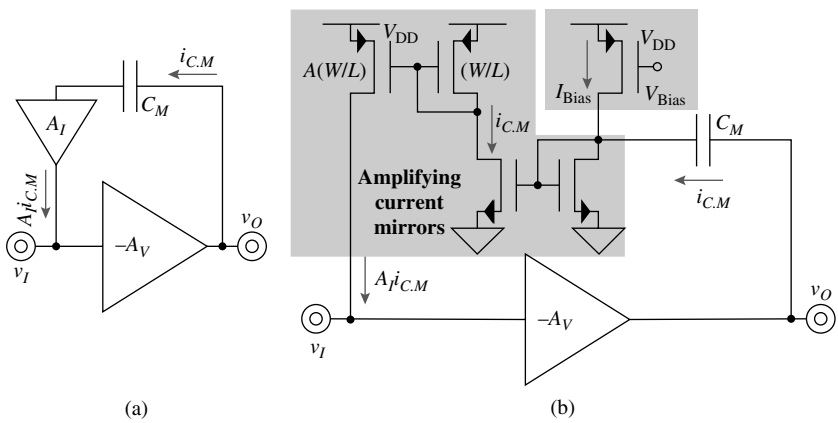


Figure 4.22 (*a*) General current-mode Miller-multiplied circuit and (*b*) a transistor-level mirrored-amplified embodiment.

and

$$
\begin{aligned}
Z_{O.\mathrm{CL}} &= \frac{R_{O.\mathrm{OL}} \parallel Z_{I.\mathrm{FB}}}{1 + A_{R.\mathrm{OL}}\beta_{\mathrm{FB}}} \\
&= \frac{R_{O.\mathrm{OL}} \parallel Z_{I.\mathrm{FB}}}{1 + [(Z_{I.\mathrm{OL}} \parallel Z_{O.\mathrm{FB}})G_{\mathrm{OL}}(R_{O.\mathrm{OL}} \parallel Z_{I.\mathrm{FB}})]\left[\left(\frac{1}{Z_{I.\mathrm{FB}}}\right)A_l\right]} \\
&= \left.\frac{R_{O.\mathrm{OL}} \parallel \left(\frac{1}{sC_M}\right)}{1 + \left(R_{I.\mathrm{OL}} \parallel \frac{1}{sC_M}\right)G_{\mathrm{OL}}\left[R_{O.\mathrm{OL}} \parallel \left(\frac{1}{sC_M}\right)\right]sC_M A_l}\right|_{f >> \frac{1}{2\pi R_{I.\mathrm{OL}} C_M}} \\
&\approx \frac{1}{A_l G_{\mathrm{OL}}}
\end{aligned}
\tag{4.54}
$$

As alluded to in the previous discussion, the amplifying current mirror shown in Fig. 4.22*b* employs two mirrors because its output (i.e., $A_i i_{C.M}$) must remain in phase with $i_{C.M}$ for negative-feedback conditions to exist. Alternatively, buffering the Miller capacitor with one inverting current mirror around a noninverting amplifier (i.e., $+A_V$) retains the negative-feedback features sought while presenting no feed-forward paths to v_O. Nevertheless, irrespective of the implementation, current mirrors are relatively benign to the Miller feedback loop because the poles they introduce normally reside at high frequencies, past the loop's unity-gain frequency. The reason for this fortuitous result is the ac nodes in current mirrors offer characteristically low resistances to ac ground so the poles the parasitic capacitors produce land at relatively high frequencies.

4.6.4 Left-Half-Plane (LHP) Zeros

Using left-half-plane (LHP) zeros to cancel secondary and/or parasitic poles, as done with the nulling Miller resistor in Fig. 4.20, is another viable, though often costly compensation strategy (in terms of power and/or silicon real estate). As stated earlier, the zero should not precede the secondary pole because it would otherwise extend the compensated unity-gain frequency $f_{0\mathrm{dB.C}}$ into the parasitic-pole region. One zero can therefore offset the effects of the secondary pole below $f_{0\mathrm{dB.C}}$ but other zeros should offset those of higher frequency poles at or above $f_{0\mathrm{dB.C}}$. Outside of the nulling Miller resistor, there are two other general means of creating a LHP zero, but only as part of a zero-pole pair combination.

Passive LHP Zeros

Figure 4.23 illustrates how a series RC load presents an additional zero-pole pair to the one already expected with any ac node. Modeling an ac node with its Norton-equivalent output resistance R_{Loop}, capacitance C_{Loop}, and transconductance G_{Loop} helps highlight the effects of an added RC load. Without the additional RC load, for instance, the pole associated with the Norton-equivalent circuit occurs at the frequency when capacitor impedance $1/sC_{Loop}$ equals R_{Loop}, at $1/2\pi R_{Loop}C_{Loop}$. However, adding an RC load as shown changes the frequency response. At low frequencies, for example, the voltage gain is as before, at $G_{Loop}R_{Loop}$, but at higher frequencies, when capacitor impedance $1/sC_{z\text{-}p}$ is considerably lower than $R_{z\text{-}p}$ (i.e., past pole p_{loop}), the gain-setting resistance drops to the parallel combination of R_{Loop} and $R_{z\text{-}p}$ (i.e., $R_{Loop} \,||\, R_{z\text{-}p}$). Leveling the gain to $G_{Loop}(R_{Loop} \,||\, R_{z\text{-}p})$ at p_{loop}, as illustrated in Fig. 4.23*b*, effectively cancels the gain-dropping effect of p_{loop} like a LHP zero would (i.e., $z_{z\text{-}p}$). Generally, a load that includes a series RC combination to ac ground introduces a LHP zero-pole-pair combination (i.e., $z_{z\text{-}p}$ and $p_{z\text{-}p}$) in the frequency response of the circuit.

From a design perspective, with respect to adding a LHP zero, assuming intentional capacitor $C_{z\text{-}p}$ is substantially larger than parasitic capacitor C_{Loop} is reasonable so the pole responsible for the drop in gain (i.e., p_{loop}) results when equivalent capacitance impedance $1/sC_{z\text{-}p}$ shunts $R_{Loop} + R_{z\text{-}p}$ at approximately $1/2\pi(R_{Loop} + R_{z\text{-}p})C_{z\text{-}p}$:

$$\left.\frac{1}{sC_{z\text{-}p}}\right|_{p_{loop}\approx\frac{1}{2\pi(R_{Loop}+R_{z\text{-}p})C_{z\text{-}p}}} \equiv R_{Loop} + R_{z\text{-}p} \tag{4.55}$$

(Refer to the time-constant technique presented in Chap. 3 for an explanation on how to derive poles by inspection.) The zero responsible for

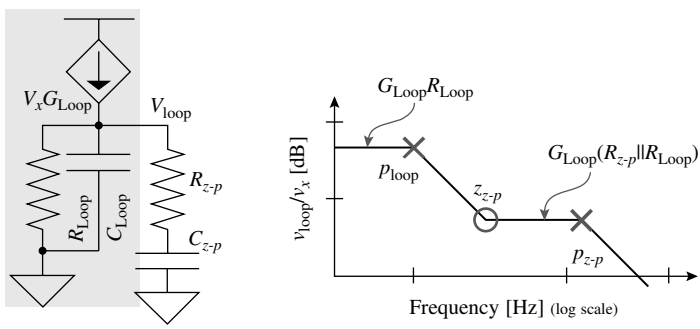


Figure 4.23 (*a*) Passive RC zero-pole pair generator and (*b*) its resulting frequency response.

leveling the gain (i.e., $z_{z\text{-}p}$) at higher frequencies occurs when $1/sC_{z\text{-}p}$ falls below $R_{z\text{-}p}$, the onset of which happens when $1/sC_{z\text{-}p}$ equals $R_{z\text{-}p}$ at about $1/2\pi R_{z\text{-}p}C_{z\text{-}p}$:

$$\left.\frac{1}{sC_{z\text{-}p}}\right|_{z_{z\text{-}p}=\frac{1}{2\pi R_{z\text{-}p}C_{z\text{-}p}}} \equiv R_{z\text{-}p} \tag{4.56}$$

At even higher frequencies, capacitor impedance $1/sC_{\text{Loop}}$ shunts the remaining impedance present (which is now $R_{z\text{-}p} \,||\, R_{\text{Loop}}$) at or above roughly $1/2\pi(R_{z\text{-}p} \,||\, R_{\text{Loop}})C_{\text{Loop}}$, introducing yet another pole to the response (i.e., $p_{z\text{-}p}$):

$$\left.\frac{1}{sC_{\text{Loop}}}\right|_{p_{z\text{-}p}\approx\frac{1}{2\pi(R_{z\text{-}p}||R_{\text{Loop}})C_{\text{Loop}}}} \equiv R_{z\text{-}p} \,||\, R_{\text{Loop}} \tag{4.57}$$

In the end, two poles and a zero result: the pole expected with any ac node (i.e., p_{loop}), albeit at a slightly modified frequency, and an additional zero-pole pair (i.e., $z_{z\text{-}p}$ and $p_{z\text{-}p}$). As such, p_{loop} can establish the dominant low-frequency pole in a negative-feedback loop and $z_{z\text{-}p}$ can cancel the secondary pole, and $p_{z\text{-}p}$ must remain at frequencies well above the compensated unity-gain frequency of the loop.

Active LHP Zeros

Inserting the same RC combination in the feedback path of an inverting op-amp configuration, as illustrated in Fig. 4.24*a*, or in its transistor-level equivalent, yields similar results to the pole-zero-pole blend shown in Fig. 4.23*b* and now shown in Fig. 4.24*b*. More explicitly, equivalent pole-zero pair p_{loop}-$z_{z\text{-}p}$ results because the closed-loop gain of the circuit is flat at $-R_2/R_1$ at low frequencies and again flat, but lower at $-(R_2 \,||\, R_{z\text{-}p})/R_1$, at frequencies higher than $1/2\pi R_{z\text{-}p}C_{z\text{-}p}$, when capacitor impedance $1/sC_{z\text{-}p}$ is equal to or smaller than $R_{z\text{-}p}$. In other words,

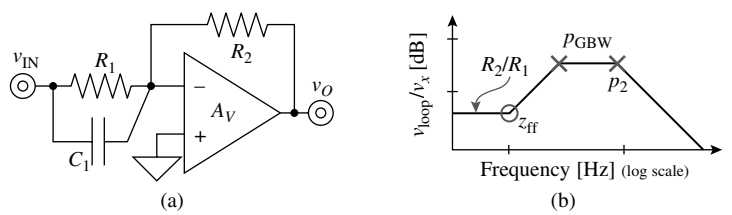


Figure 4.24 (*a*) Using a capacitor in the input impedance of an op amp to introduce a zero in (*b*) the frequency response.

$1/sC_{z\text{-}p}$ reduces the gain by reducing the equivalent feedback impedance (i.e., shorting $R_2 + R_{z\text{-}p}$) at $p_{z\text{-}p}$,

$$\left.\frac{1}{sC_{z\text{-}p}}\right|_{p_{z\text{-}p}=\frac{1}{2\pi(R_2+R_{z\text{-}p})C_{z\text{-}p}}} \equiv R_2 + R_{z\text{-}p} \tag{4.58}$$

until $1/sC_{z\text{-}p}$ becomes smaller than $R_{z\text{-}p}$ past $z_{z\text{-}p}$, beyond which frequency $1/sC_{z\text{-}p}$ is negligible, $R_{z\text{-}p}$ dominates, and gain again remains relatively constant:

$$\left.\frac{1}{sC_{z\text{-}p}}\right|_{z_{z\text{-}p}=\frac{1}{2\pi R_{z\text{-}p}C_{z\text{-}p}}} \equiv R_{z\text{-}p} \tag{4.59}$$

At even higher frequencies, the closed-loop bandwidth of the circuit introduces yet another pole to the frequency response (i.e., $p_{z\text{-}p}$ in Fig. 4.23*b*). This last pole corresponds to the loop-gain translation of the op amp's bandwidth (from negative feedback), which is equivalent to gain-bandwidth product GBW (or pole p_{GBW}) and equates to the op amp's unity-gain frequency $f_{0\text{dB}}$, if compensated to exhibit a single-pole roll-off response.

A slight and perhaps simpler adaptation of the active filter modifies the input impedance, as shown in Fig. 4.25*a*, not the feedback impedance, as in the last case, so that the input impedance changes with frequency. As in the previous example, the low-frequency gain that results is $-R_2/R_1$, but unlike before, feed-forward capacitor C_1 increases the gain when impedance $1/sC_1$ shunts R_1 at and above zero z_{FF} (i.e., $1/2\pi R_1C_1$) and levels only past the loop-gain translation of the op amp's bandwidth (i.e., p_{GBW} or unity-gain frequency $f_{0\text{dB}}$), as shown in Fig. 4.25*b*:

$$\left.\frac{1}{sC_1}\right|_{z_{\text{ff}}=\frac{1}{2\pi R_1C_1}} \equiv R_1 \tag{4.60}$$

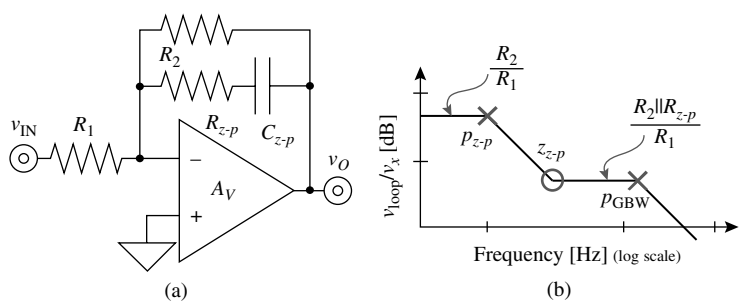


Figure 4.25 (*a*) Using an RC filter in the feedback impedance of an op amp to produce a pole-zero pair in (*b*) the frequency response.

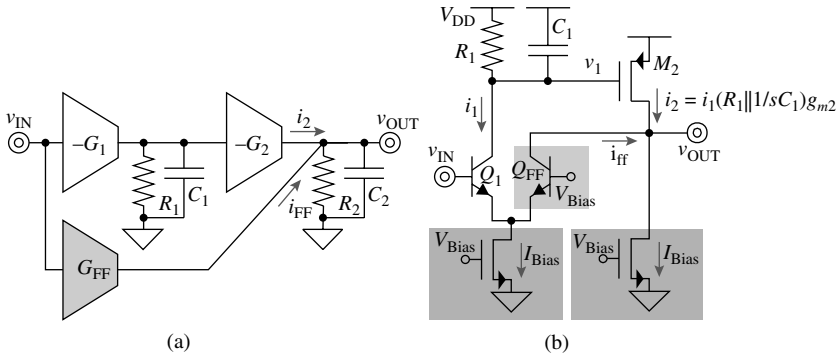


FIGURE 4.26 (*a*) Using transconductor G_{FF} to channel in-phase feed-forward signals (and introduce LHP zeros) and (*b*) a transistor-level embodiment.

Negative-feedback effects cease past the unity-gain frequency of the op amp, beyond which frequency the output is at the mercy of its loading capacitance, introducing yet another pole to the circuit (i.e., p_2 in the figure).

Just as the Miller capacitor presents a RHP zero by feed-forwarding an out-of-phase signal, an *in-phase* feed-forward path as shown in Fig. 4.26*a* introduces a LHP zero, if its magnitude is less than that of its main path at lower frequencies and higher at higher frequencies. For instance, main-signal current i_2 decreases with frequency past $1/2\pi R_1C_1$ (i.e., pole p_1), and past LHP zero z_{LHP}, when its magnitude is less than feed-forward current i_{FF}, i_{FF} dominates and sets (i.e., flattens) the transconductance gain of the circuit to a higher value (i.e., G_{FF}):

$$i_2 = \frac{v_{IN}G_1R_1G_2}{1+sR_1C_1}\bigg|_{f>>\frac{1}{2\pi R_1C_1}}$$

$$\approx \frac{v_{IN}G_1G_2}{sC_1}\bigg|_{z_{LHP}=\frac{G_1G_2}{2\pi G_{FF}C_1}} \equiv i_{FF} = v_{IN}G_{FF} \qquad (4.61)$$

where z_{LHP} is $G_1G_2/2\pi G_{FF}C_1$. As a result, low-frequency transconductance gain $G_1R_1G_2$ decreases past p_1 and levels to G_{FF} past z_{LHP}. With respect to the overall voltage gain of the circuit, capacitor impedance $1/sC_2$ shunts R_2 past output pole p_2 at $1/2\pi R_2C_2$, irrespective of p_1 and z_{LHP}, that is, before or after p_1 or z_{LHP}, depending on where p_2 resides relative to p_1 and z_{LHP}. Figure 4.26*b* illustrates a transistor-level embodiment of the same concept where one output of a differential pair constitutes the main signal-processing path (i.e., i_1-to-i_2 translation) and the other the lower gain in-phase feed-forward path (i.e., i_{ff}). As before, the LHP zero (i.e., z_{LHP}) occurs when small-signal current i_2

drops (because of $1/2\pi R_1 C_1$ or p_1) past its feed-forward counterpart i_{ff} at $g_{m1}g_{m2}/2\pi g_{mff}C_1$:

$$i_2 = \frac{0.5v_{in}g_{m1}R_1g_{m2}}{1+sR_1C_1}\Bigg|_{f>>p_1=\frac{1}{2\pi R_1C_1}}$$

$$\approx \frac{0.5v_{IN}g_{m1}g_{m2}}{sC_1}\Bigg|_{z_{LHP}=\frac{g_{m1}g_{m2}}{2\pi g_{mff}C_1}} \equiv i_{ff} = 0.5v_{in}g_{mff} \tag{4.62}$$

4.6.5 Design Objectives

In summary, the overriding objective in stabilizing a negative-feedback circuit is to pull the lowest frequency pole p_1 to sufficiently low frequencies to ensure loop gain $A_{OL}\beta_{FB}$ falls with only one pole to unity-gain frequency f_{0dB}. Keeping secondary pole p_2 near f_{0dB} trades settling time for speed (i.e., delay) and optimizes power consumption because pushing p_2 beyond f_{0dB} requires additional power. Offsetting the effects of p_2 with a left-half-plane zero (i.e., z_{LHP}) helps extend f_{0dB} (i.e., improve speed and reduce delay) without increasing settling time (i.e., compromising stability), which means p_2 and z_{LHP} can precede f_{0dB} slightly. All other parasitic poles in the circuit should remain at least one decade above f_{0dB} (i.e., $10f_{0dB}$) because, even if their individual effects on phase are small near f_{0dB}, their combined effect avalanches around $5f_{0dB}$. Naturally existing and additional left-half-plane zeros slightly above f_{0dB} mitigate these avalanching effects, extending the frequency range in which f_{0dB} can reside. Unfortunately, right-half-plane zeros not only extend f_{0dB} closer to the parasitic pole region but also reduce phase margin so altogether avoiding them is the best approach, and keeping them at least one decade above f_{0dB} the next best alternative.

Parasitic poles, in general, bound f_{0dB} so designers aim to eliminate, cancel, or shift them to out-of-reach frequencies. Eliminating poles amounts to decreasing the number of ac nodes in the feedback loop, which is why using an arbitrary number of gain stages in the feedback loop is not acceptable, as more poles compromise the stability of the circuit. The net effect of limiting the number of nodes, however, is lower loop gain. While some margin may exist to decrease the gain across a loop, the gain must remain reasonably larger than 1 for the benefits of negative feedback to apply. As such, the designer must shift the parasitic signal-shunting poles that result from increasing the loop gain to higher frequencies by either adding localized shunt-feedback loops or decreasing their associated resistances (e.g., increasing their bias currents) or capacitances (e.g., using smaller transistors).

4.7 Summary

The entire field of analog IC design may not *all* converge on negative feedback but its direct or indirect dependence to it is certainly prevalent in most complex systems today. The fact is negative feedback provides a means to control and regulate currents and/or voltages against variations in supply, loading, and environmental conditions in applications ranging from power supplies and thermostat controllers to analog drivers, filters, and data converters. Its most appealing feature is control, as the loop responds to ensure a virtual short or mirror exists between an incoming signal and a feedback counterpart, the ramifications of which are predictable gain and extremely high or low input/output impedances.

From the perspective of analysis, however, the power of negative feedback rests on its resulting approximations, as derived from its virtual short/mirror characteristics. Recognizing how the incoming signal mixes with the loop (and how the circuit samples or senses the output) is critical because the effects on impedance are vastly different. Sensing voltages or mixing currents in parallel (i.e., shunt), for instance, decreases the impedance by the circuit's loop gain while mixing voltages or sensing currents in series does the opposite. Ultimately, for all this to work, the dc gain across the loop and mixer must remain negative and considerably greater than one. A positive-feedback loop, for reference, feeds itself *in phase* and causes the output to grow and "*rail*" against the supplies, or fall into sustained oscillations. To prevent similar conditions from happening in a negative-feedback counterpart, the loop gain must not be –1 (or equivalently, 1 ∠180°) because the denominator of the closed-loop-gain expression would approach zero and the closed-loop gain peak to infinity. Losing control over its output in this way is extremely undesirable, of course, which is why including phase margin at the unity-gain frequency is so important. Note the loop gain (i.e., $A_{OL}\beta_{FB}$), and not the forward open-loop gain (i.e., A_{OL}), must be unity-gain stable.

The general strategy for compensating a negative feedback loop is to ensure the loop gain has a single-pole response, with secondary and parasitic poles at or above the compensated unity-gain frequency. There are several ways to achieve this objective, from adding a pole and using a Miller capacitor to establish a dominant low-frequency pole to canceling secondary and parasitic poles with left-half-plane zeros. Miller compensation is popular because of its pole-splitting feature, except, if unabated, it also introduces a right-half-plane zero—series impedances and/or unidirectional buffers can offset and even cancel those out-of-phase feed-forward effects. Ultimately, irrespective of the compensation strategy, the compensated unity-gain frequency is *always* below its uncompensated counterpart; and how close the former is to the latter is a measure of success, albeit at the

possible cost of power, silicon real estate, and proneness to instabilities. With respect to voltage regulators, one of the most challenging design aspects is maintaining stable operating conditions in the presence of a widely variable and unpredictable load. Chapter 5 delves deeper into the requirements such a load presents and discusses how and when to address them.

CHAPTER 5
AC Design

The most important feature of a supply circuit is its ability to regulate the output signal against dc and ac variations in its load and input supply, which is why negative feedback and regulators appear hand in hand. In voltage regulators, as illustrated in Fig. 5.1, the negative feedback loop shunt-samples output voltage v_{OUT} to ensure load-current variations have little impact on v_{OUT} and series-mixes a sensed version of v_{OUT} (via feedback voltage v_{FB}) with dc input reference voltage V_{REF} to establish a virtual short between V_{REF} and v_{FB}, directly relating and regulating v_{OUT} to V_{REF}. Finite loop gains, however, limit the efficacy of this regulating short, giving rise to nonzero *load-* (LDR) and *line-regulation* (LNR) effects on the output, that is, small variations in v_{OUT} when confronted with changes in load current and input line voltage v_{IN}.

In mobile battery-operated devices, load currents may vary by several orders of magnitude, from maybe 1–10 μA when idling to 50–150 mA during heavy loading conditions. This variation has a substantial impact on the frequency response of the regulator's output for which the circuit must remain stable. The tolerance and temperature dependence of the equivalent series resistance (ESR) associated with the output capacitor exacerbate the variability of the output filter (C_{Filter} in Fig. 5.1), forcing the designer to accommodate worst-case conditions and tradeoff loop gain and bandwidth for stability, both of which limit the regulating performance of the supply.

In most portable electronics today, the linear regulator derives its power and energy from off- or on-chip switching ac-dc or dc-dc converters whose outputs carry considerable systematic ac fluctuations at their respective switching frequencies and related harmonics. Higher quiescent current in the linear regulator affords the designer more flexibility with which to suppress this noise but the cost is, of course, higher battery drain current, in other words, shorter operational life. Higher supply current is especially problematic under light loads, which is a prevalent state in mobile applications, because quiescent current is sufficiently dominant to determine battery life. Restraining bias currents, unfortunately, constrain loop gain and bandwidth, degrading LDR and LNR performance across frequency,

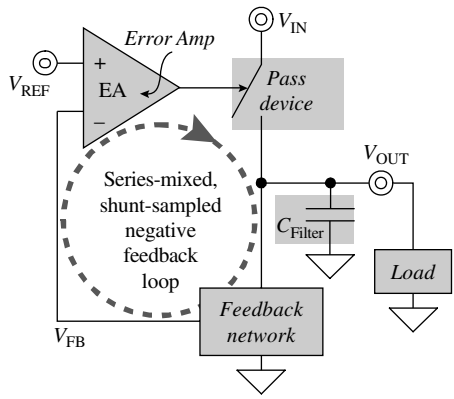


Figure 5.1 Loaded linear regulator.

especially at and above the unity-gain frequency of the regulator (i.e., f_{0dB}), when feedback is no longer effective. The objective of this chapter is to therefore study the stability requirements of a linear regulator and ascertain their impact on loop gain and bandwidth, in other words, on load-dump induced variations on v_{OUT} and *power-supply rejection* (PSR). Incidentally, LDR and LNR refer to large-signal dc performance, load-dump variations to large- and small-signal transiently induced changes in v_{OUT}, and PSR to small-signal ac effects.

5.1 Frequency Compensation

5.1.1 Uncompensated Response

Decomposing the general circuit shown in Fig. 5.1 into its component-level equivalent, as illustrated in Fig. 5.2*a*, is useful in determining the frequency-response implications of the loaded output. Although the error amplifier can ultimately take the form of one of many possible combinations of active and passive devices, its core must always embed a gain-setting stage with a characteristically high-impedance node v_A so its most basic and somewhat idealized embodiment reduces to a transconductor (e.g., G_A), resistor (e.g., R_A), and capacitor (e.g., C_A) combination. The two Norton-equivalent circuits that dependent transconductor G_P, resistor R_P, and capacitor C_P and load current i_L, resistor R_L, and capacitor C_L comprise model the effects of the pass device and load, respectively. To complete the model (and close the feedback loop), the shunt-sampling feedback network translates v_{OUT} to v_{FB} via a passive voltage divider that consists of resistors R_{FB1} and R_{FB2}.

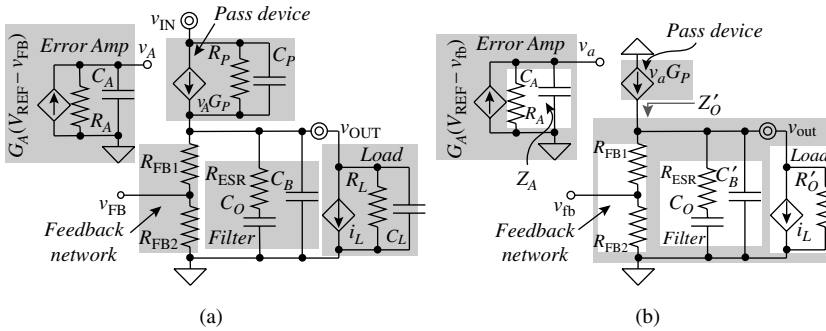


FIGURE 5.2 (*a*) Component-level translation of a loaded linear regulator and (*b*) its small-signal loop-gain equivalent.

Although not always the case, the ESR of the output filter capacitor C_O (i.e., R_{ESR}) may be large enough to induce the end user to introduce one or several high-frequency (i.e., low-ESR) bypass capacitors (i.e., C_B) near the *point of load* (PoL) with C_O-to-C_B ratios (i.e., C_O/C_B) as low as 10 µF/µF. The presence of C_B, however, is less common in system-on-chip (SoC) implementations because the real estate of the printed-circuit-board (PCB) and the costs associated with off-chip low-ESR capacitors are often prohibitively large. Nevertheless, capacitors and resistors C_P, C_L, and C_B and R_P and R_L, within the context of small-signal loop gain, are in parallel and their individual effects on frequency response are therefore analogous, which is why equivalent bypass capacitance C'_B and output resistance R'_O in Fig. 5.2*b* often absorb them for simplicity. Note that in the case of PSR analysis, however, which the next section describes, v_{in} is no longer zero, which means C_p and R_p feed-forward ac signals emanating from v_{in} to v_{out}, invalidating the C'_B and R'_O reduction.

Loop Gain

The *loop gain* (LG) of the circuit is the ac *open-loop* gain through the mixer and across the loop, that is, the error-to-feedback signal gain v_{fb}/v_e or $v_{fb}/(v_{ref} - v_{fb})$. This parameter is equivalent to the gain through the loop when the loop is open, when, for example, feedback-signal component v_{fb} in the mixer (i.e., error amplifier's transconductor) is zero. LG is therefore the product of the forward gain $G_A Z_A G_P Z'_O$, where Z_A and Z'_O represent equivalent impedances at v_a and v_{out}, respectively, and feedback ratio $R_{FB2}/(R_{FB1} + R_{FB2})$:

$$\text{LG} \equiv \frac{v_{fb}}{v_e} = \frac{v_{fb}}{v_{ref} - v_{fb}} = G_A Z_A G_P Z'_O \left(\frac{R_{FB2}}{R_{FB1} + R_{FB2}} \right) \tag{5.1}$$

The error amplifier introduces a pole p_A via Z_A when C_A shunts R_A (i.e., the impedance across C_A equals R_A), which happens at relatively low-to-moderate frequencies, at $1/2\pi R_A C_A$:

$$\left.\frac{1}{sC_A}\right|_{p_A=\frac{1}{2\pi R_A C_A}} \equiv R_A \tag{5.2}$$

To ascertain the effects of composite output impedance Z'_O on frequency response, it helps to evaluate the gain across the pass device (i.e., v_{out}/v_a or $G_P Z'_O$) at low frequencies and ascertain what happens as frequency increases. Accordingly, at low frequencies, when referring to the Z'_O block in Fig. 5.2*b* (or the simplified version shown in Fig. 5.3*a*), C_O and C'_B are open-circuited and Z'_O reduces to R''_O or the parallel combination of R'_O and $(R_{FB1} + R_{FB2})$, giving a low-frequency pass-device gain of $G_P R''_O$ or $G_P[R'_O || (R_{FB1} + R_{FB2})]$. At slightly higher frequencies, in spite impedance $1/sC_O$ decreases with frequency, the impedance across C_O continues to overwhelm R_{ESR}, which is on the order of milliohms, reducing the series combination of $1/sC_O$ and R_{ESR} to $1/sC_O$. The low-frequency gain therefore remains unaffected until C_O and C'_B shunt R''_O, when the combined impedance across C_O and C'_B equals or falls below $R'_O || (R_{FB1} + R_{FB2})$ or R''_O, at which point v_{out} shunts to ground and decreases at 20 dB per decade of frequency, introducing pole p_O in the frequency response shown in Fig. 5.3*b* at $1/2\pi R''_O(C_O + C'_B)$:

$$\left(R_{ESR} + \frac{1}{sC_O}\right) || \left(\frac{1}{sC'_B}\right)\Bigg|_{\text{Low Freq.}} \approx \left.\frac{1}{s(C_O + C'_B)}\right|_{p_O=\frac{1}{2\pi R''_O(C_O+C'_B)}}$$

$$\equiv R'_O || (R_{FB1} + R_{FB2}) = R''_O \tag{5.3}$$

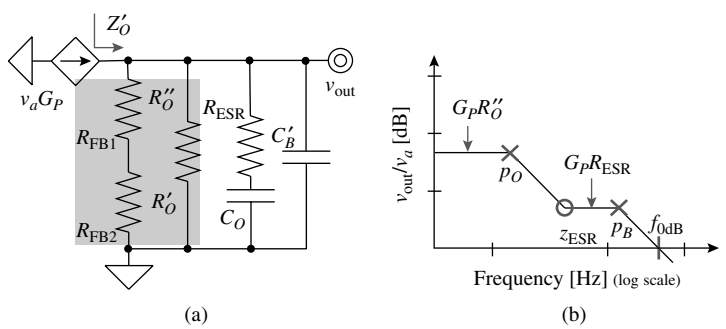


Figure 5.3 (*a*) Simplified small-signal model of the pass device and (*b*) its resulting gain.

At higher frequencies, when $1/sC_O$ falls substantially below R_{ESR}, the series combination of R_{ESR} and $1/sC_O$ reduces to R_{ESR}, ending the effects of C_O on v_{OUT}. Because C_B' is considerably smaller than C_O, impedance $1/sC_B'$ is not yet a short circuit and Z_O' consequently reduces to $R_O'' \,||\, R_{ESR} \,||\, (1/sC_B')$ or roughly R_{ESR}, as $1/sC_B'$ is also larger than R_{ESR} at moderate frequencies. This behavior yields a flat moderate-frequency pass-device gain of $G_P R_{ESR}$ at and above zero z_{ESR}, when impedance $1/sC_O$ equals R_{ESR}, at $1/2\pi C_O R_{ESR}$:

$$\left.\frac{1}{sC_O}\right|_{z_{ESR}=\frac{1}{2\pi R_{ESR}C_O}} \equiv R_{ESR} \tag{5.4}$$

At these (and higher) frequencies, G_P, R_{ESR}, and C_B' determine the gain across the switch and C_O and R_O'' no longer incur noticeable effects. As frequencies further increase, C_B' shunts v_{out} when the impedance across C_B' falls below R_{ESR}, which happens past pole p_B, at $1/2\pi R_{ESR}C_B'$:

$$\left.\frac{1}{sC_B'}\right|_{p_B=\frac{1}{2\pi R_{ESR}C_B'}} \equiv R_{ESR} \tag{5.5}$$

Deriving the full expression for Z_O' algebraically yields

$$\begin{aligned} Z_O' &= \frac{R_O''(1+sR_{ESR}C_O)}{s^2R_O''R_{ESR}C_OC_B' + s(R_O'' + R_{ESR})C_O + sR_O''C_B' + 1} \\ &\approx \frac{R_O''(1+sR_{ESR}C_O)}{s^2R_O''R_{ESR}C_OC_B' + sR_O''(C_O + C_B') + 1} \end{aligned} \tag{5.6}$$

The numerator in Z_O' carries the aforementioned zero (i.e., z_{ESR}) explicitly and the denominator decomposes into output and bypass poles p_O and p_B. Note that extracting p_O and p_B from this more explicit expression is easier when realizing R_{ESR} is substantially smaller than R_O'' and C_O larger than C_B'; in other words, $R_O'' + R_{ESR}$ and $C_O + C_B'$ reduce to R_O'' and C_O, respectively. As a result, because the s^2 term is negligibly small at low frequencies (i.e., s^2 term can be discarded) and R_O'' is substantially higher than R_{ESR} (i.e., $R_O'' + R_{ESR}$ reduces to R_O''), the denominator gives rise to p_O. Similarly, the s^2 and s terms are considerably larger than 1 at high frequencies so discarding the s^0 term (i.e., 1) in the denominator, factoring s, and using C_O for $C_O + C_B'$ exposes p_B. In summary, as seen in Fig. 5.3, Z_O' introduces two poles (i.e., p_O and p_B) and one zero (i.e., z_{ESR}) to the loop gain of the regulator.

Ultimately, the overall loop gain of a linear regulator has three poles and one zero: error amplifier pole p_A, output pole p_O, bypass

pole p_B, and ESR zero z_{ESR}. Referring to Fig. 5.2*b*, the low-frequency loop gain (i.e., LG_{LF}) is

$$\begin{aligned} LG_{LF} &= G_A R_A G_P R_O'' \beta_{FB} \\ &= G_A R_A G_P [R_L \,||\, R_P \,||\, (R_{FB1} + R_{FB1})] \left(\frac{R_{FB2}}{R_{FB1} + R_{FB2}} \right) \end{aligned} \tag{5.7}$$

and the overall loop gain across frequency is

$$\begin{aligned} LG &\approx \frac{LG_{LF}\left(1 + \dfrac{s}{2\pi z_{ESR}}\right)}{\left(1 + \dfrac{s}{2\pi p_A}\right)\left(1 + \dfrac{s}{2\pi p_O}\right)\left(1 + \dfrac{s}{2\pi p_B}\right)} \\ &\approx \frac{G_A R_A G_P [R_L \,\|\, R_P \,\|\, (R_{FB1} + R_{FB2})]\left(\dfrac{R_{FB2}}{R_{FB1} + R_{FB2}} \right)\{1 + sR_{ESR}C_O\}}{\{1 + sR_A C_A\}\{1 + [R_L \,\|R_P\|\, (R_{FB1} + R_{FB2})](C_O + C_B + C_L + C_P)\}\{1 + sR_{ESR}(C_B + C_L + C_P)\}} \end{aligned} \tag{5.8}$$

where $R_L \,||\, R_P$, $R_L \,||\, R_P \,||\, (R_{FB1} + R_{FB2})$, and $C_B + C_L + C_P$ comprise R_O', R_O'', and C_B', respectively. Figure 5.4 illustrates the generalized, though uncompensated loop-gain response of a linear regulator. Note C_O suppresses transient-induced output voltage fluctuations, which is why C_O is typically large, anywhere from 1 nF for low-power regulators to more than 10 µF for higher power applications, placing output pole p_O at relatively low frequencies. High loop gains are similarly desirable because the ability of the circuit to regulate v_O against load and line variations is directly proportional to this gain. As a result, for

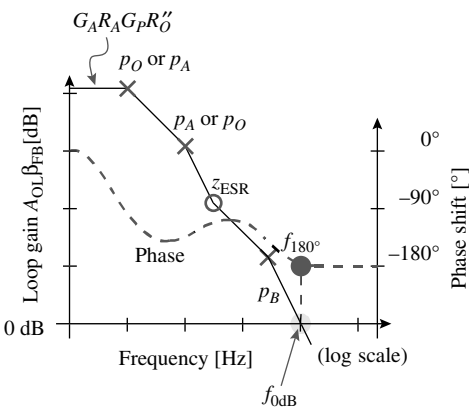


Figure 5.4 Uncompensated loop-gain Bode-plot response of a linear regulator.

moderate gains, the error amplifier normally requires a high-impedance node that necessarily places a pole (i.e., p_A) at relatively lower frequencies. Which of these two poles precedes the other depends on the circuit, and more specifically, on the type of power device used.

At first glance, ESR zero z_{ESR} seems to save phase, but in practice, its location is highly unpredictable and therefore unreliable. The fact is ESR resistor R_{ESR} varies considerably across process and temperature and characterization data is generally poor, the latter of which implies simulation models are correspondingly deficient. As a result, z_{ESR} may or may not lie within the bandwidth of the regulator, and intentionally adding series resistance to C_O to keep z_{ESR} within band only degrades transient-response and PSR performance, as partially discussed in Chap. 1 and later addressed here, in this chapter. The same variation applies to bypass pole p_B because it also depends on R_{ESR}, which explains why p_B tracks z_{ESR} with respect to variations in R_{ESR}. Pole p_B is always at higher frequencies than z_{ESR} by a p_B/z_{ESR} ratio of approximately $C_O/(C_B + C_P + C_L)$ or C_O/C'_B; in other words, p_B is the product of C_O/C'_B and z_{ESR}.

In general, relatively high-power breadboards and printed-circuit-boards (PCBs) often include z_{ESR} and p_B within the bandwidth of the regulator because high-power loads are physically displaced from C_O and attached to low-ESR capacitors for better transient and PSR response, forcing the designer to use high bypass capacitances. Conversely, lower power system-in-package (SiP) and system-on-chip (SoC) solutions place the regulator near their respective, but now lower power loads, at the point of load (PoL). The proximity to relatively lower power loads often negates the need for bypass capacitors and exorbitantly high output capacitances, the latter of which may afford the designer the flexibility of using high-frequency output capacitors, effectively reducing R_{ESR} to the point where its effects appear well above frequencies of interest. Eliminating z_{ESR} and p_B, however, still leaves a potentially unstable two-pole response, and that is excluding the parasitic poles the error amplifier introduces, which are often difficult to place at high frequencies under low-power constraints.

5.1.2 Externally Compensated Response

The dominant low-frequency pole of the *externally compensated* circuit, by definition, as applied to this textbook, is at the output of the regulator. To keep output pole p_O dominant, it is convenient to use large output capacitors and p-type pass transistors in common-source configurations because they offer higher RC time constants. In the case of a PMOS device, for instance, transconductance, resistance, and capacitance G_P, R_P, and C_P are small-signal transconductance, drain-source resistance, and drain-bulk capacitance g_m, r_{ds}, and C_{DB}. The aspect ratio of the power transistor is typically large with as short a channel length as the lithography and breakdown limits of the process allow because short-channel lengths drive higher currents while introducing lower parasitic capacitances, in other words, because short-channel-length

devices are faster. As a result, PMOS channel-length modulation parameter λ can be as high as 0.1 V^{-1}, producing output resistances around 0.1–10 kΩ when confronted with 1–100 mA loads (i.e., r_{ds} is roughly $1/\lambda I_L$, where I_L is the steady-state load current and the quiescent current flowing through the feedback network is negligible when compared to I_L).

These small r_{ds} resistances are typically well below feedback and load resistances $R_{FB1} + R_{FB2}$ and R_L, reducing the low-frequency output resistance and output pole (i.e., p_O) to r_{ds} and

$$p_O \approx \frac{1}{2\pi r_{ds}(C_O + C_B + C_L + C_P)} \approx \frac{I_L}{2\pi(C_O + C_B + C_L + C_{DB})} \tag{5.9}$$

respectively, with the latter being directly proportional to steady-state load current I_L. Because I_L normally spans four decades or more of steady-state current variation (e.g., from 5 μA to 50 mA), p_O shifts by an equally expansive range, as illustrated in Fig. 5.5*a*. Although the gain across the pass device at low frequencies (i.e., A_P, which is roughly $G_P r_{ds}$) and output pole p_O are directly and inversely proportional to r_{ds}, the gain-bandwidth product of the loop (i.e., GBW_{PMOS}, which is proportional to $A_P p_O$) and, by translation, the resulting unity-gain frequency (i.e., $f_{0dB.MOS}$, which is proportional to GBW_{PMOS}) do not escape the effects of a variable load. To be specific, A_P's dependence on the transconductance of the pass device (i.e., G_P) allows variations in I_L to shift $f_{0dB.MOS}$ with the square root of those changes (i.e., $f_{0dB.MOS}$ is directly proportional to $\Delta I_L^{1/2}$):G

$$GBW_{PMOS} \propto A_P p_O \propto \frac{g_{m.PMOS} r_{ds}}{r_{ds}} = \sqrt{2 I_L \left(\frac{W}{L}\right) K'_P} \propto \sqrt{I_L} \tag{5.10}$$

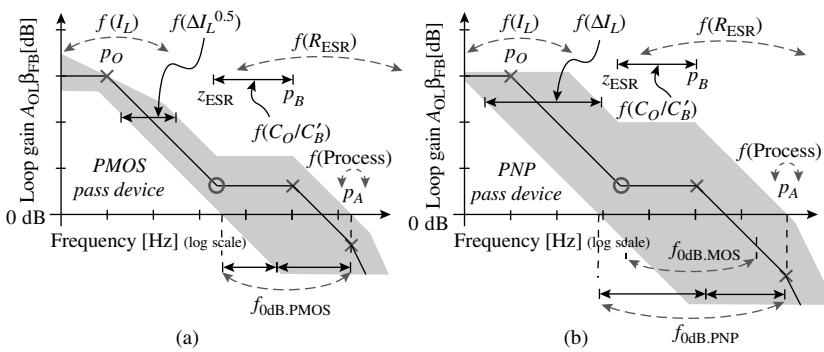


Figure 5.5 Frequency-response variation space for externally compensated (*a*) common-source (CS) PMOS and (*b*) common-emitter (CE) PNP regulators.

Unlike the PMOS case, the transconductance of a PNP device is roughly I_L/V_t, in other words, linearly proportional to I_L so its effects on gain-bandwidth product GBW_{PNP} and unity-gain frequency $f_{0\text{dB.PNP}}$ are more significant, shifting them with ΔI_L instead of $\Delta I_L^{1/2}$, as illustrated in Fig. 5.5*b*:

$$\text{GBW}_{\text{PNP}} \propto A_P p_O \propto \frac{g_{m.\text{PNP}} r_o}{r_o} = \frac{I_L}{V_t} \propto I_L \tag{5.11}$$

The pole variation of the PNP case mimics the PMOS version because r_o (as r_{ds}) is inversely proportional to I_L:

$$p_O \approx \frac{1}{2\pi r_o (C_O + C_B + C_L + C_C)} \approx \frac{I_L}{2\pi V_A (C_O + C_B + C_L + C_C)} \tag{5.12}$$

where C_C is the parasitic collector-substrate capacitance of the transistor. The low-frequency gain across the PNP transistor, however, unlike the MOS case, is independent of I_L because $g_{m.\text{PNP}}$'s direct linear dependence to I_L cancels r_o's inverse linear dependence: $g_{m.\text{PNP}} r_o$ is $(I_L/V_t)(V_A/I_L)$ or V_A/V_t.

Equivalent series resistance (ESR) variations further expand the frequency range where unity-gain frequency $f_{0\text{dB}}$ resides. On one extreme, ESR resistor R_{ESR} may be sufficiently small (at maybe less than 25–50 mΩ) to push z_{ESR} and p_B well above $f_{0\text{dB}}$ and allow the gain to continually drop at 20 dB per decade past p_O and yield a relatively low $f_{0\text{dB}}$. On the other extreme, R_{ESR} may be large enough (like for instance, greater than 50–500 mΩ) to pull z_{ESR} and p_B within frequencies of interest and extend $f_{0\text{dB}}$ by the frequency ratio of z_{ESR} and bypass pole p_B (i.e., extend $f_{0\text{dB}}$ by a factor of C_O/C'_B), as seen in Fig. 5.5. Pole p_A must therefore reside at or above this highest possible worst-case frequency location of $f_{0\text{dB}}$. Ultimately, $f_{0\text{dB}}$'s net maximum-minimum ratio (i.e., $f_{0\text{dB(max)}}/f_{0\text{dB(min)}}$) represents the compounded effects of variations in I_L on the gain-bandwidth product at lower frequencies and the possible presence of zero-pole pair z_{ESR}-p_B on f_{0dB} at moderate frequencies:

$$\frac{f_{0\text{dB(max).PMOS}}}{f_{0\text{dB(min).PMOS}}} = \left(\frac{\text{GBW}_{(\text{max})}}{\text{GBW}_{(\text{min})}}\right)\left(\frac{p_{B(\text{max})}}{z_{\text{ESR(min)}}}\right) \approx \left(\sqrt{\frac{I_{L(\text{max})}}{I_{L(\text{min})}}}\right)\left(\frac{C_{O(\text{max})}}{C'_{B(\text{min})}}\right) \tag{5.13}$$

and

$$\frac{f_{0\text{dB(max).PNP}}}{f_{0\text{dB(min).PNP}}} = \left(\frac{\text{GBW}_{(\text{max})}}{\text{GBW}_{(\text{min})}}\right)\left(\frac{p_{B(\text{max})}}{z_{\text{ESR(min)}}}\right) \approx \left(\frac{I_{L(\text{max})}}{I_{L(\text{min})}}\right)\left(\frac{C_{O(\text{max})}}{C'_{B(\text{min})}}\right) \tag{5.14}$$

for the PMOS and PNP cases, respectively. The maximum-minimum unity-gain-frequency ratios resulting from a load that demands 1–100 mA with 1 µF and 0.1 µF of output and bypass capacitances, for example, are 100 and 1,000 Hz/Hz, respectively, for the PMOS and PNP cases.

The highly variable frequency response just described presents several challenges. For one, the error amplifier must comprise low-resistance and low-capacitance nodes to produce poles (e.g., p_A and others) that reside at relatively high frequencies. This condition is difficult to satisfy when considering the amplifier must also yield high gain for load- and line-regulation performance, for which large resistances are generally necessary, and drive the large parasitic capacitance the power pass device presents. Second, guaranteeing a high unity-gain frequency for reasonable load-dump and supply ripple-rejection performance amounts to increasing low-frequency loop gain LG_{LF} and pushing p_A and all relevant parasitic poles to even higher frequencies, requiring more quiescent current, producing lower power efficiency, and shortening single-charge battery lifetime. These design constraints under the low-quiescent current requirements typically attached to battery-powered solutions force the lowest worst-case loop gain and unity-gain frequency to remain at approximately 40–50 dB and 10–100 kHz.

In optimizing loop gain, bandwidth, and quiescent power, several worst-case corner conditions may result. The lowest worst-case f_{0dB} location occurs when z_{ESR} and p_B are high (i.e., R_{ESR} is low), p_O is low (i.e., C_O, C_B, C_P, and C_L are high and I_L is low), and the gain across the pass device is low, which happens at the weakest process and temperature corner of the pass device. One worst-case stability condition usually results when f_{0dB} extends into the parasitic-pole region, which corresponds to low z_{ESR} frequencies (i.e., high R_{ESR} values), high p_B frequencies (i.e., low C_B, C_L, and C_P values), high p_O frequencies (i.e., low C_O and high I_L values), and high gains (which occur at the strongest process and temperature corners of the pass device). The other worst-case stability scenario results when the ESR zero resides one or more decades above f_{0dB}, where it no longer saves phase, which occurs with low R_{ESR} values, leaving a possibly unstable two-pole system, if output and error-amplifier poles p_O and p_A reside at higher frequencies (with, for instance, high I_L and low C_O, C_B, C_P, and C_L values). The highest low-frequency loop gain possible corresponds to the first worst-case stability condition because higher loop gains further extend f_{0dB}, which means loop gain must remain below an upper bound (e.g., less than 55–60 dB) to prevent f_{0dB} from creeping to higher frequencies. Generally, the tradeoff corners of a PNP pass device may be more difficult to manage because of its increased gain-bandwidth product and resulting f_{0dB} variations with respect to I_L.

5.1.3 Internally Compensated Response

The dominant low-frequency pole of *internally compensated* regulators is inside the negative-feedback loop, that is to say, not at the output,

as in the externally compensated counterparts. To keep output pole p_O at higher frequencies, it is convenient to use smaller output capacitors and n-type follower pass transistors because they offer lower RC time constants. As discussed in Chap. 3, the gain across these devices is roughly unity and relatively insensitive to load current I_L. As a result, because the error amplifier's gain is also relatively independent of I_L, the gain-bandwidth product of the loop GBW_N and its corresponding unity-gain frequency $f_{0\text{dB}.N}$ are relatively insensitive to variations in load current I_L, as illustrated in Fig. 5.6*a*:

$$\text{GBW}_N \propto A_P p_A = \frac{\left(\dfrac{g_m R_O''}{1+g_m R_O''}\right)}{2\pi R_A C_A} \approx \frac{1}{2\pi R_A C_A} \neq I_L \tag{5.15}$$

Miller-compensated p-type transistors are also useful in internally compensated regulators because they introduce a gain-setting, dominant low-frequency pole that resides inside the loop (as p_A) while simultaneously pushing output pole p_O to higher frequencies. As in the follower case, gain-bandwidth product $\text{GBW}_{\text{Miller}}$ and unity-gain frequency $f_{0\text{dB.Miller}}$ are independent of I_L, as shown in Fig. 5.6*b*, because pole-setting capacitor C_A is, for all practical purposes, directly proportional to the gain across the p-type transistor, assuming Miller-compensating capacitor C_C is considerably larger than all other parasitic capacitors:

$$\text{GBW}_{\text{Miller}} \propto A_P p_A = \frac{g_m R_O''}{2\pi R_A C_A} \approx \frac{g_m R_O''}{2\pi R_A (g_m R_O'' C_C)} = \frac{1}{2\pi R_A C_C} \neq I_L \tag{5.16}$$

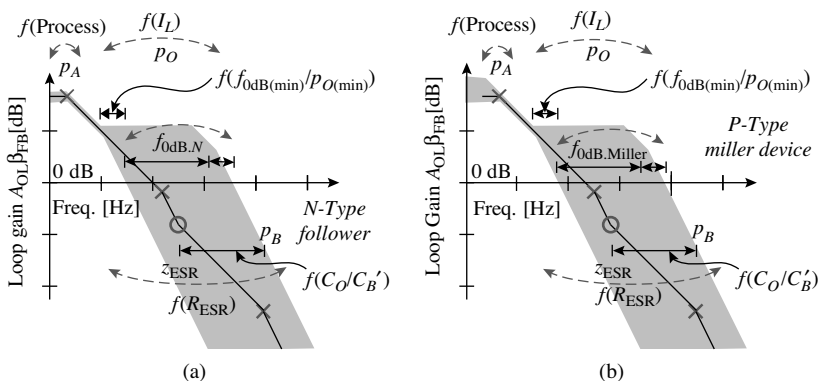


Figure 5.6 Frequency-response variation space for internally compensated (*a*) n-type common-collector/drain (CC/CD) follower and (*b*) p-type common-emitter/source (CE/CS) Miller regulators.

Output pole p_O depends on output resistance $1/g_m$ in both the follower and Miller-compensated cases, and because large output capacitors are desirable (for transient performance), its location may nominally lie below the unity-gain frequency of the circuit,

$$p_O \approx \frac{g_m}{2\pi(C_O + C_B')} = \frac{g_m}{2\pi(C_O + C_B + C_P + C_L)} \tag{5.17}$$

Again, because load current I_L traverses several orders of magnitudes and R_{ESR} spans a wide range of values across process and temperature, g_m and consequently p_O, z_{ESR}, and p_B vary substantially, which is why p_O, under extreme conditions, may precede both z_{ESR} and unity-gain frequency f_{0dB}. As in the externally compensated regulator, the presence of z_{ESR} extends f_{0dB} and variations in I_L, through changes in p_O, expand its range. Note there is no change in gain to offset changes in p_O and p_O does not always lie below f_{0dB}. The lowest possible p_O frequency must be high enough, in the absence of z_{ESR}, to assure stable conditions for the negative-feedback loop, limiting the extent p_O expands f_{0dB}'s variation space to less than a decade, for placing p_O a decade or more below f_{0dB} decreases phase margin to prohibitive levels (e.g., to zero):

$$\frac{f_{0dB.Int(max)}}{f_{0dB.Int(min)}} = \left(\frac{f_{0dB.Int}}{p_{O(min)}}\right)\left(\frac{p_{B(max)}}{z_{ESR(min)}}\right) \le 10\left(\frac{C_{O(max)}}{C_{B(min)} + C_{P(min)} + C_{L(min)}}\right) \tag{5.18}$$

where $f_{0dB.Int}$ generally refers to the unity-gain frequencies of internally compensated regulators.

Again, the highly variable frequency response just described presents a number of design challenges. In this case, output pole p_O must remain at relatively high frequencies, which is difficult because output capacitor C_O is necessarily large to mitigate the impact of fast load dumps on output v_{OUT}. Maintaining high bandwidth is also difficult when considering light-to-zero loading conditions pull p_O to considerably lower frequencies, closer to p_A, the worst-case result of which arises when z_{ESR} is above f_{0dB} (when R_{ESR} is low) and unable to save phase. These constraints typically limit the output capacitance of the circuit to relatively lower values, curbing the regulator's ability to suppress high-power load-dumps variations in v_{OUT} and therefore constraining the circuit to lower power applications, where load dumps are more manageable.

In optimizing loop gain, bandwidth, and quiescent power, similar worst-case-corner conditions to the externally compensated regulator result. The lowest worst-case unity-gain frequency (i.e., f_{0dB}), for instance, occurs when z_{ESR} is absent (i.e., R_{ESR} is low) and not able to extend f_{0dB} to high frequencies, p_O is low enough (i.e., C_O is high and I_L low) to pull f_{0dB} to lower frequencies, and the gain across the pass

device is low (which occurs at the pass device's weakest process and temperature corner). The same conditions, with the exception of high gain (i.e., high z_{ESR}, low p_O, and the strongest process and temperature corner), may also produce a worst-case stability condition because no zero exists to salvage the phase shift induced by p_A and p_O. Another worst-case stability situation may result when f_{0dB} extends into the parasitic-pole region, when z_{ESR} is low (with high R_{ESR} values), p_B high (with low C_B, C_L, and C_P values), p_O high (with low C_O and high I_L values), and gain high (at the strongest process and temperature corner). The lowest loop gain corresponds to the second worst-case stability case, except at the low-gain process and temperature corner, because a higher loop gain would otherwise extend f_{0dB} in the strongest corner into the parasitic-pole region, where stability may be compromised.

5.2 Power-Supply Rejection

Power-supply rejection (PSR), or as is also commonly known, *power-supply ripple rejection* or *ripple rejection*, for short, refers broadly to the ability of a circuit to regulate its output against low- and high-frequency small-signal (i.e., ac) variations in the input supply (i.e., line voltage). PSR is therefore defined as the complement of supply *injection*, or equivalently, the reciprocal of supply gain A_{IN}, the latter of which is the small-signal variation in output voltage ∂v_{OUT} or v_{out} that results in response to small-signal changes in input supply v_{IN} (i.e., ∂v_{IN} or v_{in}); in other words, A_{IN} is the ac supply-output gain v_{out}/v_{in}:

$$\text{PSR} \equiv \frac{1}{A_{\text{IN}}} = \frac{\partial v_{\text{IN}}}{\partial v_{\text{OUT}}} \equiv \frac{v_{\text{in}}}{v_{\text{out}}} \tag{5.19}$$

Similarly, but with respect to steady-state signals, LNR is supply gain A_{IN} at low frequencies (i.e., at steady state), in other words, dc variations in v_{OUT} or ΔV_{OUT} in response to steady-state changes in v_{IN} or ΔV_{IN}. Because LNR applies to the linear region of the regulator, which means large-signal changes in v_{IN} have a linear effect on v_{OUT}, just as small-signal v_{in} has on v_{out}, LNR is the complement or reciprocal of PSR at low frequencies (i.e., PSR_{LF}), which is equivalent to low-frequency supply gain $A_{IN.LF}$:

$$\text{LNR} \equiv \frac{\Delta V_{\text{OUT}}}{\Delta V_{\text{IN}}} \approx \left.\frac{\partial v_{\text{OUT}}}{\partial v_{\text{IN}}}\right|_{\text{LowFreq.}} = A_{\text{IN.LF}} = \frac{1}{\text{PSR}_{\text{LF}}} \tag{5.20}$$

It is unfortunate that "power-supply ripple rejection" produces the same acronym as "*power-supply rejection ratio*" (PSRR), which designers most often apply to amplifiers, not regulators. The two terms are related and similar but not the same. The latter, for instance,

refers to how much a circuit favors changes in input signal over variations in the supply so the ratio of input-output gain A_V to supply-output gain A_{DD} sets PSRR:

$$\text{PSRR} \equiv \frac{A_V}{A_{DD}} \tag{5.21}$$

unlike the former, which refers to how much a circuit is able to reject supply ripples—the reciprocal or inverse of supply-output gain A_{IN} (or equivalently A_{IN}^{-1} or A_{DD}^{-1}). Applying a supply ripple to an operational amplifier in unity-gain configuration, however, which resembles a linear regulator setup, is a rather useful means of extrapolating the PSRR performance of an operational amplifier because, as A_{DD} transfers ripple v_{dd} to v_{out}, unity-gain feedback produces a loop-gain translation of the same at v_{out}, which roughly equals PSRR:

$$v_{out} = v_{dd}A_{DD} + v_{id}A_V = v_{dd}A_{DD} - v_{out}A_V \tag{5.22}$$

or

$$\frac{v_{dd}}{v_{out}} = \frac{A_{DD}}{1 + A_V\beta_{FB}}\bigg|_{\beta_{FB}=1} = \frac{A_{DD}}{1 + A_V} \approx \frac{A_{DD}}{A_V} = \frac{1}{\text{PSRR}} \tag{5.23}$$

This approximation is only valid, with respect to PSRR, as long as forward gain A_V is considerably larger than 1 V/V, in other words, up to unity-gain frequency f_{0dB}, which is where PSRR differs from PSR, because the latter is accurate past f_{0dB}. Ultimately, to avoid confusion, this book only uses acronym PSR to refer to ripple-rejection performance.

5.2.1 Shunt-Feedback Model

Since PSR is nothing more than the inverse of supply-output gain A_{IN} or v_{out}/v_{in}, it helps to decompose the linear regulator into its voltage-divider (or impedance-ladder) equivalent. Series feedback-network resistance $R_{FB1} + R_{FB2}$, load-impedance combination $R_L \,||\, (1/sC_L)$, and filter capacitors, for instance, as illustrated in Fig. 5.7*a* and *b*, comprise lower ladder impedance Z_O, where dc load current I_L is an open circuit for small-signal analysis. Impedance Z_O differs from Z_O' (as discussed earlier) in that the latter, through C_B' and R_O' (and R_O''), includes pass capacitance C_P and resistance R_P and only applies to loop-gain analysis with respect to frequency compensation, where v_{IN} is an ac ground (i.e., zero). Supply voltage v_{IN}, in this case, is an ac source and the pass device consequently offers a direct feed-through impedance path to v_{OUT} through R_P and C_P (i.e., via Z_P).

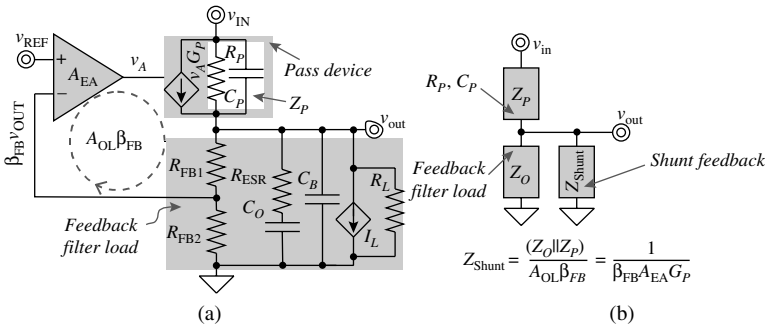


FIGURE 5.7 (*a*) Linear regulator and (*b*) its power-supply rejection (PSR) small-signal voltage-divider equivalent.

The transconductance of the pass device (i.e., G_p) is the means through which shunt-sampled feedback asserts its effects. It may also carry supply-dependent signals, as with a p-type pass device where the output current is proportional to the difference between v_{IN} and shunt-feedback signal v_A, but not always, as in the case of n-type followers. Generally, though, the transconductor always carries shunt-feedback signal v_A. Neglecting, for the time being, v_{IN} components in transconductor G_p whose effects will be addressed shortly after this discussion, the equivalent impedance G_p presents from v_{OUT} to ground (i.e., Z_{Shunt}), as shown in Fig. 5.7*b*, is the shunt-feedback-reduced version of open-loop output impedance $Z_P \| Z_O$ (i.e., Z_{Shunt} is $(Z_P \| Z_O)/A_{OL}\beta_{FB}$). Since loop gain $A_{OL}\beta_{FB}$ is itself a function of $Z_P \| Z_O$, however, Z_{Shunt} reduces to the reciprocal of loop-gain transconductance i_{GP}/v_{out} (i.e., G_{FB}) or $1/\beta_{FB}A_{EA}G_P$, where A_{EA} is the gain across the error amplifier and β_{FB} the feedback factor through R_{FB1} and R_{FB2}:

$$Z_{Shunt} \equiv \frac{v_T}{i_T} \equiv \frac{v_{out}}{i_{GP}} = \frac{1}{G_{GP}} = \frac{v_{out}}{v_{out}\beta_{FB}A_{EA}G_P} = \frac{1}{\beta_{FB}A_{EA}G_P} \tag{5.24}$$

or

$$Z_{Shunt} = \frac{1}{\beta_{FB}A_{EA}G_P} = \frac{Z_P \| Z_O}{\beta_{FB}[A_{EA}G_P(Z_P \| Z_O)]} = \frac{Z_P \| Z_O}{\beta_{FB}A_{OL}} \tag{5.25}$$

where v_T and i_T are generic ac test signals used to determine the impedance from a particular point to ac ground. This result corroborates general shunt-feedback analysis whose closed-loop output impedance $Z_{O.CL}$ is the loop-gain-reduced translation of open-loop

output impedance $Z_{O.OL}$ in parallel with the input impedance of the feedback network $Z_{I.FB}$, or simply $Z_O \| Z_P$:

$$Z_{O.CL} = Z_{Shunt} \| Z_O \| Z_P = \frac{Z_{Shunt}(Z_O \| Z_P)}{Z_{Shunt}(Z_O \| Z_P)} = \frac{Z_O \| Z_P}{1 + A_{OL}\beta_{FB}} = \frac{Z_{O.OL} \| Z_{I.FP}}{1 + A_{OL}\beta_{FB}} \tag{5.26}$$

The model presented in Fig. 5.7*b* also makes intuitive sense in that shunt feedback presents a *shunting* impedance (i.e., Z_{Shunt}) to v_{OUT}, that is to say, to Z_O.

5.2.2 Feed-through Components in G_P

Common-Mode Concept

As mentioned in the previous discussion, transconductor G_P in the pass device may also feed v_{in}-dependent components to v_{out}, which are, of course, undesirable in linear regulators. These adverse feed-through effects depend on how the pass transistor and error amplifier relate at the circuit level. In considering n-type pass transistors, for instance, as shown in Fig. 5.8*a* and *b*, the n-type followers, for all practical purposes, impress and replicate whatever v_{in}-dependent ac signals are present in v_A onto v_{OUT}. As a result, within the context of linear regulators, no v_{in}-dependent ac components should exist in v_A when using n-type switches (Fig. 5.8*b*). In other words, gate-source and base-emitter terminals should be common mode with respect to each other—in this case, if the gate or base terminal does not carry noise, neither does the corresponding source or emitter.

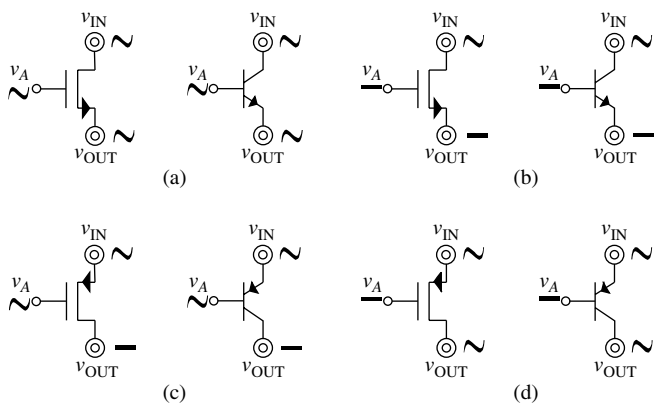


Figure 5.8 Illustration of the effects of common-mode signals on output v_{OUT} in (*a*)-(*b*) n-type and (*c*)-(*d*) p-type transistors.

P-type pass devices present an entirely different scenario. To eliminate v_{IN}-dependent ac signals (i.e., v_{in}) from feeding through the transistor to v_{OUT}, feedback signal v_A should carry the same v_{IN}-dependent ac signal (i.e., v_a should be v_{in}) so that the resulting transconductor current i_o and therefore small-signal output voltage and supply gain $A_{IN.P}$ are zero with respect to v_{in}:

$$A_{IN.P(v_a=v_{in})} = \frac{v_{out}}{v_{in}} = \frac{i_o Z_O}{v_{in}} = \frac{(v_{in} - v_a) g_m Z_O}{v_{in}}\bigg|_{v_a=v_{in}} = 0 \qquad (5.27)$$

That is to say, gate or base signal v_A must be *common mode* with respect to its corresponding source or emitter, which carries v_{in} in this case. If on the other hand, v_A has no ac dependence to v_{IN}, the p-type pass device becomes a common-gate (CG) current buffer amplifier to v_{IN}, producing a nonzero ac transconductor current i_o and consequently a nonzero and in-phase ohmic voltage drop across the output and a nonzero $A_{IN.P}$ value:

$$A_{IN.P(v_a=0)} = \frac{v_{out}}{v_{in}} = \frac{i_o Z_O}{v_{in}} = \frac{(v_{in} - v_a) g_m Z_O}{v_{in}}\bigg|_{v_a=0} = g_m Z_O \qquad (5.28)$$

Generally, as with n-type devices, gate and base terminals should be common mode with respect to their source and emitter terminals, which is another way of saying p-type gates and bases should carry the same noise present at their respective sources and emitters: v_{in}.

Current Mirrors

Current mirrors, as expected, replicate all v_{IN}-dependent ac currents they receive at their inputs in their outputs. Ground-referenced (n-type) current mirrors, for instance, as shown in Fig. 5.9*a*, sink whatever ac currents flow into their inputs (e.g., $i_{in} = v_{in}/R$) at their outputs via their respective output transconductance sources (e.g., $i_{gm} \approx i_{in} \approx v_{in}/R$), assuming the resistance of the diode-connected input transistor (i.e., $1/g_m$) is substantially smaller than any other resistance in its vicinity. Similarly, supply-referenced (p-type) mirrors (Fig. 5.9*b*) source whatever ac currents their inputs conduct through their respective output transconductance sources.

The most significant conclusion to draw from these observations, from the perspective of PSR, is the direction, or equivalently, the polarity of the ac signal that appears at the output. Ground-referenced mirrors, for example, subtract v_{IN}-dependent ac components from their outputs while their supply-referenced counterparts do the opposite: add. This fact is useful in creating common-mode signals, as will be discussed shortly.

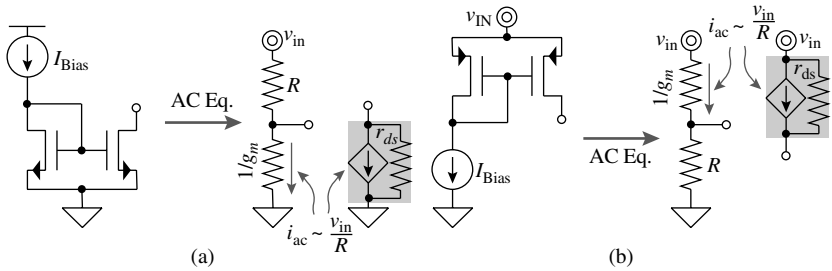


Figure 5.9 Supply ac signals in (*a*) ground- and (*b*) supply-referenced current mirrors.

Ground-Referenced Differential Amplifiers

What is perhaps most important within the context of linear regulators is how error-amplifier voltage v_A relates to v_{IN}; in other words, what the error amplifier's supply gain A_{IN} (or v_a/v_{in}) is. Figure 5.10 illustrates the schematic and small-signal translation of a differential amplifier (e.g., MD1-MD2) loaded with a ground-referenced mirror (e.g., MM1-MM2). Since the shunt-feedback model presented earlier already accounts for feedback effects in the output of the regulator, the ac input signals of the amplifier, for the foregoing analysis, are 0 V. Because both inputs are at the same bias potential (e.g., V_{BiasD}), differential pair MD1-MD2 decomposes into its common-mode half-circuit equivalent so twice the tail current resistance (i.e., $2r_{dsTail}$) degenerates each side of the differential pair with respect to small signal v_{in}. Replacing the degenerated transistors by their degenerated-equivalent resistances R_D, which roughly equate to $2g_{mD}r_{dsD}r_{dsTail}$ (i.e., $R_D = g_{mD}r_{dsD}2r_{dsTail} + r_{dsD} + 2r_{dsTail} \approx 2g_{mD}r_{dsD}r_{dsTail}$), further reduces the circuit to a voltage divider-mirror combination.

The reduced model of the ground-referenced amplifier mimics the current-mirror circuit analyzed earlier, except the output is now loaded with a resistor equivalent to the one producing the v_{in}-dependent

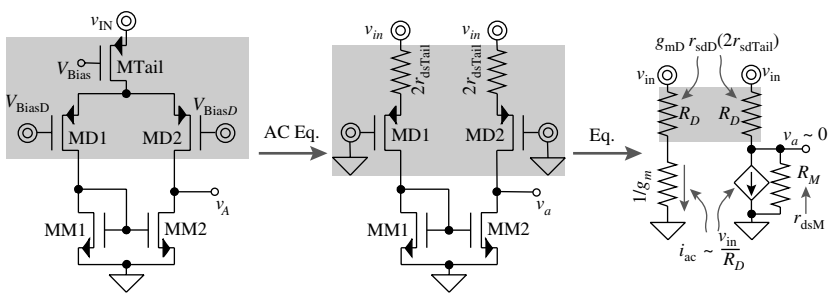


Figure 5.10 P-type differential pair with ground-referenced (n-type) mirror and its small-signal translations.

current in the input. Employing superposition to determine the effects of dependent current source i_{ac} and independent voltage source v_{in} on output v_a, reveals the v_{in}-dependent current sank (or subtracted) by the mirroring transconductor (i.e., i_{ac} or v_{in}/R_D) cancels the positive (supplied) contribution of v_{in} in the output leg of the circuit:

$$A_{IN.A(GND.Ref)} \equiv \frac{v_a}{v_{in}} = \frac{\left(\frac{v_{in}R_M}{R_D+R_M}\right) - [i_{ac}(R_D \parallel R_M)]}{v_{in}}$$
$$\approx \frac{\left(\frac{v_{in}R_M}{R_D+R_M}\right) - \left[\left(\frac{v_{in}}{R_D}\right)(R_D \parallel R_M)\right]}{v_{in}} = 0 \qquad (5.29)$$

which means v_a, for all practical purposes, has no ac contribution from v_{in}, irrespective of the actual and relative resistances at both terminals of the differential pair (i.e., R_D) and the output of the mirror (i.e., R_M).

Referring to the pass device of the linear regulator and the common-mode conclusions drawn from Fig. 5.8, a ground-referenced amplifier, since its output (i.e., v_a) is independent of v_{in}, is good for n-type follower pass devices but *poor* for their p-type counterparts because the amplifier does not present a common-mode signal with respect to v_{IN}. Note this result is an approximation based on the relative resistance ratio of loading resistance R_D to diode-connected resistance $1/g_{mM1}$, assuming the latter is substantially smaller than the former:

$$i_{ac} = \frac{v_{in}}{R_D + \frac{1}{g_{mM1}}} \approx \frac{v_{in}}{R_D} \qquad (5.30)$$

Avoiding this assumption produces a nonzero A_{IN} value whose effects on the output of the regulator are normally negligible when compared to other feed-through contributions, especially when R_D is substantially high, as in the case just described where R_D is a cascoded resistance equivalent to approximately $2g_{mD}r_{dsD}r_{dsTail}$:

$$A_{IN.A(GND.Ref)} \equiv \frac{v_a}{v_{in}} = \frac{\left(\frac{v_{in}R_M}{R_D+R_M}\right) - i_{ac}(R_D \parallel R_M)}{v_{in}} = \left(\frac{R_M}{R_D+R_M}\right) - \left(\frac{R_D \parallel R_M}{R_D + \frac{1}{g_{mM1}}}\right) \qquad (5.31)$$

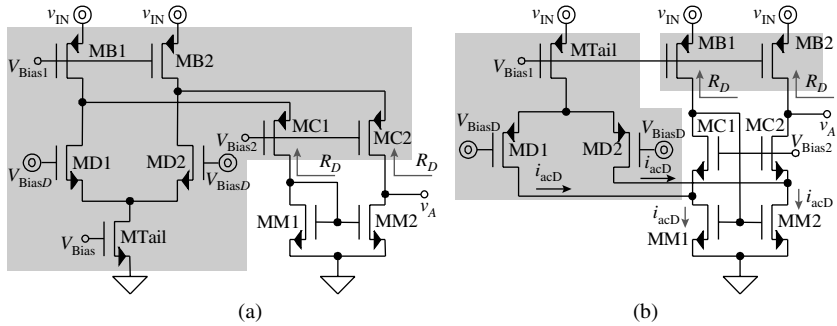


FIGURE 5.11 (*a*) N- and (*b*) p-type differential pairs with ground-referenced (n-type) load mirrors.

The symmetrical nature of the differential stage is the fundamental reason why the v_{IN}-dependent ac currents roughly cancel in the output. It is therefore not surprising that folded-cascode translations of the differential pair with ground-referenced (n-type) load mirrors also produce analogous results. The folded-cascode n-type input amplifier shown in Fig. 5.11*a*, for instance, presents symmetrical n-type MM1-MM2 loading conditions to n-type differential pair MD1-MD2 (i.e., the resistances into both MM1-MM2 drains are roughly equal and equivalent to R_D in Fig. 5.10) so current-mirror output transconductor current i_{gmM2} cancels the v_{in}-dependent variations resulting from the ac current sourced by cascode transistor MC2.

Folded-cascode amplifiers with p-type inputs and ground-referenced (n-type) load mirrors, as illustrated in Fig. 5.11*b*, also yield similar results. First, biasing transistors MB1-MB2 present symmetrical loading conditions to cascoded current mirror MM1-MM2-MC1-MC2 so their v_{IN}-induced effects cancel, as in the simpler five-transistor amplifier case just discussed. P-type differential pair MD1-MD2 now sources equal v_{IN}-dependent ac currents (i.e., i_{acD}) to both sides of the mirror because the resistance into the sources of cascode devices MC1-MC2 is considerably low at roughly $1/g_m$:

$$i_{acD} = \frac{v_{in}}{2g_{mD}r_{sdD}r_{sdTail} + \dfrac{1}{g_m}} \approx \frac{v_{in}}{2g_{mD}r_{sdD}r_{sdTail}} \tag{5.32}$$

Common-gate current buffer MC1 therefore channels the ac current from MD1 to the input of the cascoded mirror (i.e., gate of diode-connected MM1). As a result, the mirror's output transconductor "mirrors" and sinks an equal amount of current (i.e., $i_{gmM2} = i_{acD}$), steering all the ac current from MD2 to ground and not allowing any v_{in}-dependent current derived from MD1-MD2 to reach the output via common-gate current buffer MC2. In other words, mirror MM1-MM2 cancels v_{IN}-derived ripples injected by

differential pair MD1-MD2 much in the same way it cancels supply noise from bias MB1-MB2.

Supply-Referenced Differential Amplifiers

The same analytical process applies to supply-referenced differential amplifiers, a general embodiment of which Fig. 5.12 shows, except the mirror's output transconductor current i_{gmM2} and the v_{IN}-dependent ac contribution of the mirror's output resistance r_{dsM2} now have an aggregate effect on output v_a, as opposed to the canceling effect seen earlier. To start, as before, the ac input signals of the amplifier are zero because the shunt-feedback model already accounts for feedback effects. Differential pair MD1-MD2 therefore decomposes into two common-source transistors each degenerated with twice the tail current resistance (i.e., $2r_{dsTail}$) with respect to ground. Replacing the degenerated transistors with their degenerated-equivalent resistances R_D (which roughly equal $2g_{mD}r_{dsD}r_{dsTail}$) further reduces the circuit to a voltage divider-mirror combination. Employing superposition on dependent current source i_{ac} and independent voltage source v_{in} with respect to v_a reveals that the v_{in}-dependent current sourced (i.e., added) by the mirroring transconductor (i.e., i_{ac} which is approximately v_{in}/R_D) has, as does the v_{in} contribution in the output leg of the circuit, a positive effect on v_a:

$$A_{IN.A(Vin.Ref)} \equiv \frac{v_a}{v_{in}} = \frac{\left(\frac{v_{in}R_D}{R_D + R_M}\right) + i_{ac}(R_D \| R_M)}{v_{in}}$$

$$\approx \frac{\left(\frac{v_{in}R_D}{R_D + R_M}\right) + \left(\frac{v_{in}}{R_D}\right)(R_D \| R_M)}{v_{in}} = 1 \qquad (5.33)$$

reproducing the ac signal present in v_{IN} at v_a (i.e., v_a is roughly v_{in}), irrespective of the actual and relative resistances at both terminals of

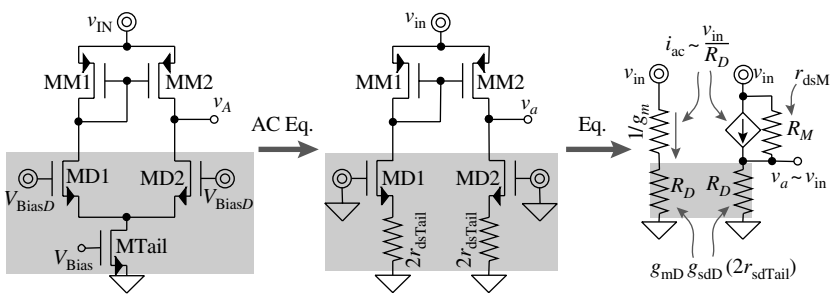


Figure 5.12 N-type differential pair with supply-referenced (p-type) mirror and its small-signal translations.

the differential pair (i.e., R_D) and the output of the mirror (i.e., R_M). Strictly speaking, however, the ac current flowing into the mirror is not exactly equal to v_{in}/R_D but R_D is so much larger than $1/g_{mM1}$ in practice that the approximation yields, for all practical purposes, the same result as the more accurate expression:

$$A_{IN.A(Vin.Ref)} \equiv \frac{v_a}{v_{in}} = \frac{\left(\frac{v_{in}R_D}{R_D + R_M}\right) + i_{ac}(R_D \,||\, R_M)}{v_{in}} = \left(\frac{R_D}{R_D + R_M}\right) + \left(\frac{R_D \,||\, R_M}{R_D + \frac{1}{g_{mM1}}}\right) \tag{5.34}$$

In referring back to the pass device of the linear regulator and the common-mode conclusions drawn from Fig. 5.8, a supply-referenced amplifier, since its output v_a also carries v_{in} signals, is therefore a *good design choice for p-type pass devices* because gate-source and base-emitter terminals are in common mode. Conversely, a supply-referenced amplifier is a poor design choice for their n-type follower counterparts. Just as the folded translations of the ground-referenced loads mimicked their five-transistor predecessor, the symmetry of the differential stage also ensures the folded-cascode translations of the supply-referenced loads have an aggregate effect on output v_a, approximately reproducing v_{in} in v_a. The folded-cascode p-type input amplifier shown in Fig. 5.13*a*, for example, presents symmetrical loading conditions (i.e., R_D) to p-type load mirror MM1-MM2 so current-mirror output transconductor current i_{gmM2} and the v_{in}-dependent variations resulting from the ac current flowing through output resistance r_{dsM2} sum and reproduce ac signal v_{in} at v_a.

Folded-cascode amplifiers with n-type inputs and supply-referenced (p-type) load mirrors, as shown in Fig. 5.13*b*, also yield

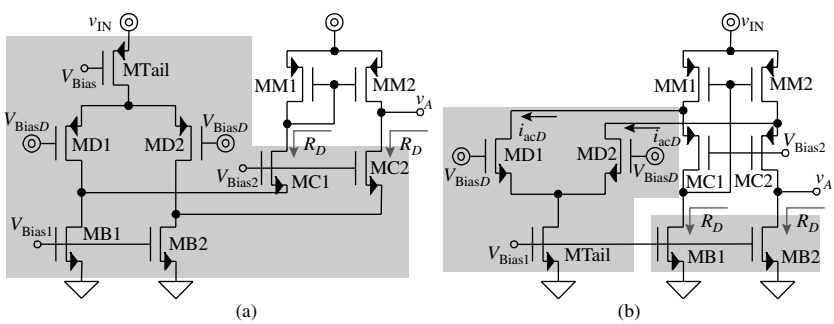


Figure 5.13 (*a*) P- and (*b*) n-type differential pairs with supply-referenced (p-type) load mirrors.

similar aggregate results. Biasing transistors MB1-MB2 present symmetrical loading conditions to cascoded current mirror MM1-MM2-MC1-MC2 so the mirror and voltage divider have a cumulative effect on v_a, yielding v_{in}-equivalent ac signals. Incidentally, differential transistors MD1 and MD2 present substantially larger resistances to ac ground (equivalent to r_{ds}) than cascode devices MC1 and MC2 (which are at $1/g_m$), assuming the loading effect of MB2 on MC2's ($1/g_m$)-source resistance is not significant. This relative resistance relationship means all ac ripple current sourced by mirror MM1-MM2 flows through MC1 and MC2, as though MD1-MD2 were absent (i.e., i_{acD} is roughly 0 A). Note that cascoding MB1 and MB2 would increase the loading effect on MC2's source resistance $1/g_{mC2}$ and establish an impedance asymmetry in the circuit that allows MD2 to steer supply ripple current away from v_a. As a result, MD2 would decrease supply injection from MC2 into v_A while MD1 would not do the same to MC1, which means the cascodes ultimately produce a ripple at the output that may not be equivalent to v_{in}.

5.2.3 Power-Supply Rejection Analysis

The v_{IN}-dependent feed-through components in the transconducntace of the pass device (i.e., G_p), as concluded in the previous subsection, hinge on the type of load mirror used in the error amplifier with respect to the pass transistor. To be more specific, feeding a feedback signal (i.e., v_A) to the base or gate of the pass device that is equivalently in phase with the emitter or source, which results from loading the error amplifier's differential pair with one of the ground- or supply-referenced mirrors shown in Figs. 5.10, 5.11, 5.12, and 5.13, eliminates feed-through effects in G_P. Given the flexibility the sample configurations shown illustrate, dwarfing the feed-forward effects of v_{IN} in transconductor G_P is relatively straightforward, simplifying the PSR ac-equivalent circuit to the Z_P-Z_O-Z_{Shunt} voltage divider shown in Fig. 5.7*b*.

In the foregoing model, pass-device impedance Z_P is the parallel combination of R_P and C_P and output impedance Z_O the parallel combination of feedback resistors $R_{FB1} + R_{FB2}$, output capacitor C_O and its ESR R_{ESR}, bypass capacitor C_B, and load R_L and C_L. Shunt-feedback impedance Z_{Shunt} (which is equivalent to $1/G_{FB}$ or $1/\beta_{FB}A_{EA}G_P$) models the supply-independent shunting effects of negative feedback on the output. Generally, and this applies to any circuit, good PSR performance results when the impedance to the supply (i.e., Z_P) and ac ground (i.e., $Z_O \,||\, Z_{Shunt}$) are as high and low as possible, respectively, which is why PSR performance benefits from the use of negative feedback (i.e., Z_{Shunt}) and bypass capacitors (i.e., C_B), because they generally decrease the impedance to ground.

Low-Frequency PSR (and LNR^{-1})

As with other ac circuits, it is often easier, and perhaps more physically germane, to determine the frequency response of the circuit

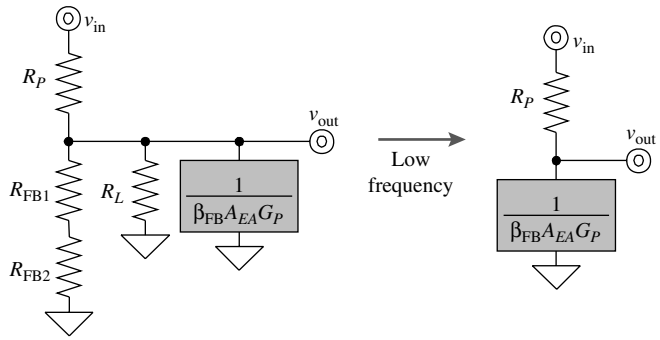


Figure 5.14 Equivalent small-signal circuit model for analyzing power-supply rejection (PSR) at low frequencies.

at low frequencies first and then ascertain what happens as frequency increases and capacitors short. Taking this approach, Z_O and Z_P decompose into $(R_{FB1} + R_{FB2}) \,||\, R_L$ and R_P, respectively, and Z_{Shunt}, which is $1/G_{FB}$, expands to $1/\beta_{FB}A_{EA}G_P$. Since Z_{Shunt} is loop-gain-reduced impedance $(Z_O \,||\, Z_P)/A_{OL}\beta_{FB}$ and low-frequency loop gain LG_{LF} or $A_{OL}\beta_{FB}$ is considerably large, Z_{Shunt} is substantially low. As a result, the parallel combination of Z_O and Z_{Shunt} simplifies to Z_{Shunt}, as shown in Fig. 5.14, and the series combination of Z_P and Z_{Shunt} (when determining the voltage-divided gain from v_{in} to v_{out}) reduces to Z_P. The low-frequency supply gain that results (i.e., $A_{IN.LF}$), which is equivalent to line-regulation performance (LNR) and the reciprocal of low-frequency power-supply rejection $(PSR)_{LF}$, is roughly inversely proportional to LG_{LF}:

$$A_{IN.LF} \equiv \frac{1}{PSR_{LF}} = \frac{v_{out}}{v_{in}}\bigg|_{Low.Freq.} = \frac{Z_O \,||\, Z_{Shunt}}{Z_P + Z_O \,||\, Z_{Shunt}}\bigg|_{Low.Freq.}$$

$$\approx \frac{Z_{Shunt}}{Z_P + Z_{Shunt}}\bigg|_{Low.Freq.} \approx \frac{Z_{Shunt}}{Z_P} = \frac{1}{(\beta_{FB}A_{EA.LF}G_P)R_P} \approx \frac{1}{LG_{LF}} \approx LNR \tag{5.35}$$

where $A_{EA.LF}$ is the error amplifier's low-frequency gain. In reality, low-frequency loop gain LG_{LF} is slightly smaller than $\beta_{FB}A_{EA}G_PR_P$ because equivalent output resistance $R_P \,||\, R_L \,||\, (R_{FB1} + R_{FB2})$ is smaller than R_P, although not by much. The reasons why the difference between R_P and $R_P \,||\, R_L \,||\, (R_{FB1} + R_{FB2})$ is often small are (1) power pass devices carry vast load currents that result in correspondingly low $1/g_m$ and r_o resistances and (2) power devices are large and built in such a way that induces more channel-length

and base-width modulation (e.g., channel-length modulation parameter λ can be as high as 0.1 V^{-1} and Early voltage V_A as low as 10 V).

Externally Compensated Response

The first parameter to have an impact on the circuit as frequencies increase in an externally compensated regulator is output pole p_O, when output, bypass, and load capacitors C_O, C_B, and C_L start to short (Fig. 5.15*a* and *b*). As a result, output impedance Z_O decreases, except its effects on PSR are at first negligible because the parallel combination of Z_O and Z_{Shunt} remains unchanged, given Z_{Shunt} is considerably low to begin with, as also demonstrated with the simulation results shown in Fig. 5.16. Although the location of error-amplifier pole p_A relative to ESR zero z_{ESR} and bypass pole p_B varies with design, if p_A precedes z_{ESR} and p_B, p_A decreases feedback transconductance G_{FB}, which increases Z_{Shunt} (Fig. 5.15*c*) and results in higher supply gain A_{IN} (i.e., lower supply injection), increasing A_{IN} at a rate of 20 dB per decade of frequency.

Impedance Z_{Shunt} (i.e., G_{FB}^{-1}) increases past p_A until Z_O, which at these frequencies decreases as C_O, C_B, and C_L continue to short to ground (because Z_O is approximately $1/s(C_O + C_B + C_L)$), is lower at

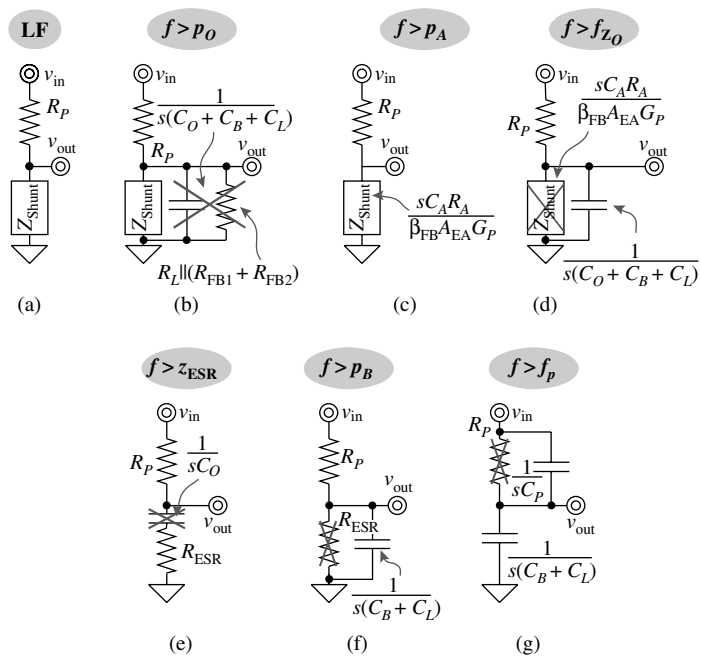


Figure 5.15 Equivalent PSR circuits as frequency increases from (*a*) low frequency (LF) past (*b*) p_O, (*c*) p_A, (*d*) f_{z_O}, (*e*) z_{ESR}, (*f*) p_B, and (*g*) f_P.

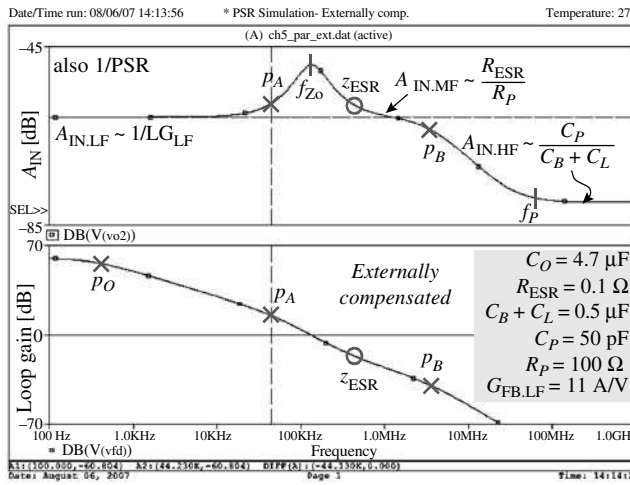


FIGURE 5.16 Loop-gain and corresponding supply-gain (i.e., A_{IN} or PSR^{-1}) simulation of an externally compensated regulator.

and past f_{Z_o} (Fig. 5.15*d*), reducing ground impedance from $Z_{Shunt} \,||\, Z_O$ to Z_O:

$$Z_O\big|_{p_O} < f < z_{ESR} \approx \left.\frac{1}{s(C_O + C_B + C_L)}\right|_{f_{Zo} \approx \frac{1}{2\pi}\sqrt{\frac{G_{FB}(2\pi p_A)}{(C_O + C_B + C_L)}}} \equiv Z_{Sh}$$

$$\approx \frac{1}{\beta_{FB}\left(\frac{A_{EA.LF} p_A}{s}\right) G_P} = \frac{s}{G_{FB.LF} p_A} \tag{5.36}$$

where $G_{FB.LF}$ is the low-frequency gain of feedback transconductance G_{FB}. The PSR voltage-divider model therefore simplifies to R_p for Z_P and $1/s(C_O + C_B + C_L)$ for $Z_O \,||\, Z_{Shunt}$, producing a supply-gain performance past f_{Zo} and before z_{ESR} (where z_{ESR} is $1/2\pi R_{ESR} C_O$) that is equivalent to $1/s(C_O + C_B + C_L)R_P$:

$$A_{IN}\big|_{f_{Zo} < f < z_{ESR}} = \left.\frac{Z_O \,||\, Z_{Shunt}}{Z_P + Z_O \,||\, Z_{Shunt}}\right|_{f_{Zo} < f < z_{ESR}}$$

$$\approx \frac{Z_O}{R_P} \approx \frac{1}{s(C_O + C_B + C_L)R_P} \tag{5.37}$$

At this point, Z_O determines PSR performance, which is why A_{IN} flattens (i.e., stops decreasing or improving) at z_{ESR}, when C_O shorts and presents constant resistance R_{ESR} from v_{out} to ground (Fig. 5.15*e*).

The voltage divider in this flat region comprises R_P (for Z_A) and R_{ESR} (for $Z_O \,||\, Z_{Shunt}$) and the resulting supply injection at moderate frequencies (MF) reduces to roughly R_{ESR}/R_P:

$$A_{IN.MF} = \left.\frac{Z_O \,||\, Z_{Shunt}}{Z_P + Z_O \,||\, Z_{Shunt}}\right|_{z_{ESR}<f<p_B} \approx \frac{Z_O}{R_P} \approx \frac{R_{ESR}}{R_P} \tag{5.38}$$

A_{IN} again drops (i.e., PSR improves) at 20 dB per decade past bypass pole p_B, when C_B and C_L start to shunt R_{ESR} (Fig. 5.15*f*).

This decreasing trend stops at higher frequencies, however, at frequency f_P, when C_P in Z_P shorts and bypasses R_P (Fig. 5.15*g*), feeding through more v_{in} signals to v_{out}:

$$\left.\frac{1}{sC_P}\right|_{f_P \approx \frac{1}{2\pi R_P C_P}} \equiv R_P \tag{5.39}$$

Note C_P only affects the circuit at high frequencies because it is parasitic and considerably smaller than C_B and C_L combined. As a result, past f_P, Z_P reduces to $1/sC_P$ and overwhelms series-impedance Z_O, and because C_B and C_L alone set impedance Z_O at those frequencies (i.e., R_{ESR} in C_O is larger), the resulting impedance divider yields a flat A_{IN} gain at approximately $C_P/(C_B + C_L)$:

$$A_{IN.HF} = \left.\frac{Z_O \,||\, Z_{Shunt}}{Z_P + Z_O \,||\, Z_{Shunt}}\right|_{f>p_B} \approx \left.\frac{Z_O}{Z_P}\right|_{f>f_P} \approx \frac{\dfrac{1}{s(C_B + C_L)}}{\dfrac{1}{sC_P}} = \frac{C_P}{C_B + C_L} \tag{5.40}$$

The foregoing analysis highlights several key points. Generally, a v_{OUT}-sensed negative-feedback loop presents a ripple-shunting impedance (i.e., Z_{Shunt}) at v_{OUT} with a value that is inversely proportional to feedback transconductance G_{FB} (i.e., Z_{Shunt} is $1/\beta_{FB} A_{EA} G_P$) and whose effects are prevalent at low-to-moderate frequencies, when Z_{Shunt} is smaller than Z_O. Error-amplifier bandwidth p_A and output, bypass, and load capacitors C_O, C_B, and C_L in output impedance Z_O subsequently produce a peak in the supply-gain response near the unity-gain frequency, at f_{Z_o}. Resistance R_{ESR} and feed-forward capacitance C_P, because they increase the noise content in v_{out}, ultimately flatten and limit supply injection at moderate-to-high frequencies (i.e., $A_{IN.MF}$ and $A_{IN.HF}$). This general behavior indicates the following:

- Low error-amplifier bandwidth p_A increases A_{IN} and therefore degrades PSR at moderate frequencies.
- High C_O, C_B, and C_L values dampen the peaking effects of low p_A's on A_{IN}; that is, they improve PSR.

- Low R_{ESR} values reduce the peak in A_{IN} at moderate frequencies, also improving PSR.
- Low C_P values keeps feed-through ripples from increasing A_{IN} at high frequencies, again improving PSR.

Internally Compensated Response

A similar analysis applies to internally compensated regulators, except the sequence of events differs because internal pole p_A precedes output pole p_O and output capacitor C_O is a low-ESR capacitor, introducing an ESR zero that characteristically resides at relatively higher frequencies. Given these differences, the first parameter to have an impact on the circuit as frequencies increase is error-amplifier bandwidth p_A. To be more specific, feedback transconductance G_{FB} decreases at and past p_A, increasing feedback impedance Z_{Shunt} (Fig. 5.17*a* and *b*) and, in consequence, increasing supply injection (i.e., degrading PSR) at 20 dB per decade, as shown in the simulation results of Fig. 5.18. For reference, assuming a single-pole response, which is to say output pole p_O is past unity-gain frequency f_{0dB} to ensure stable operating conditions, f_{0dB} is the gain-bandwidth product (GBW) of low-frequency loop gain LG_{LF} or roughly $\beta_{FB}G_AR_AG_PR_P$ and bandwidth p_A or $1/2\pi R_AC_A$:

$$f_{0dB} = \text{GBW} = \text{LG}_{\text{LF}} p_A = \frac{\beta_{FB} A_{EA} G_P Z'_O}{2\pi R_A R_A}$$
$$= \frac{\beta_{FB}(G_A R_A) G_P (Z_O \,||\, Z_P)}{2\pi C_A R_A} \approx \frac{\beta_{FB} G_A G_P R_P}{2\pi C_A} \tag{5.41}$$

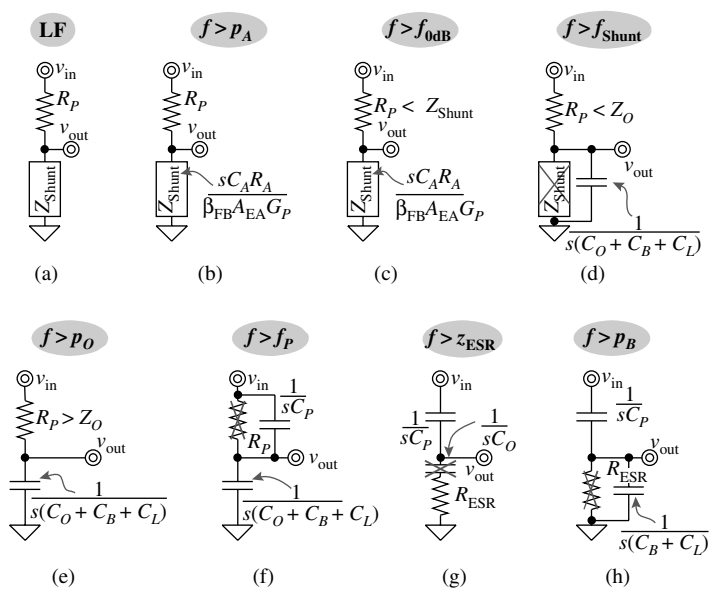


Figure 5.17 Equivalent PSR circuits as frequency increases from (*a*) low frequency (LF) past (*b*) p_A, (*c*) f_{0dB}, (*d*) p_O, (*e*) f_P, (*f*) z_{ESR}, and (*g*) p_B.

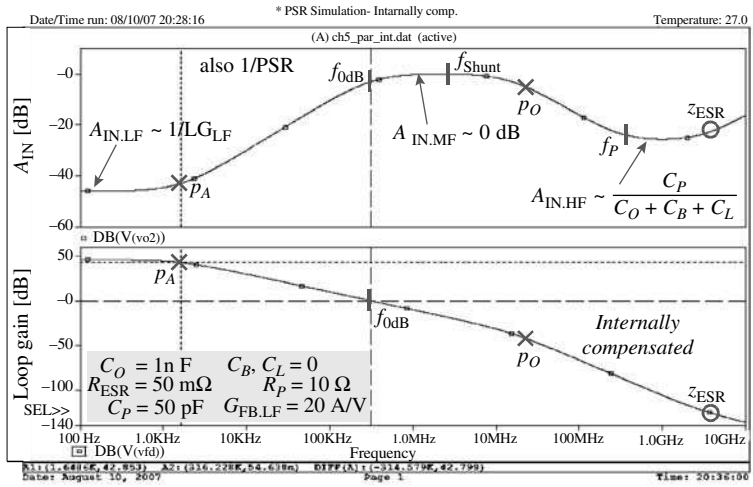


FIGURE 5.18 Loop-gain and corresponding supply-gain (i.e., A_{IN} or PSR^{-1}) simulation of an internally compensated regulator.

Past error-amplifier bandwidth p_A, supply gain continues to increase until $Z_{Shunt} \,||\, Z_O$ (which reduces to Z_{Shunt} at low-to-moderate frequencies) increases above Z_P (which in this region is simply R_P), as shown in Fig. 5.17*c*, at which point A_{IN} and PSR reach 1 V/V, or equivalently, 0 dB:

$$A_{IN.MF} = \left.\frac{Z_O \,||\, Z_{Shunt}}{Z_P + Z_O \,||\, Z_{Shunt}}\right|_{f_{0dB} < f < p_O} \approx \frac{Z_{Shunt}}{R_P + Z_{Shunt}} \approx \frac{Z_{Shunt}}{Z_{Shunt}} = 1 \qquad (5.42)$$

where $A_{IN.MF}$ refers to A_{IN} at moderate frequencies. Note Z_O at these frequencies is the parallel combination of feedback and load resistances R_{FB1} + R_{FB2} and R_L, which is considerably larger than Z_{Shunt} because p_O's shunting effects on Z_O appear at higher frequencies in internally compensated circuits, and Z_P is R_P because C_P's impedance is significantly larger than R_P. As a result, $Z_O \,||\, Z_{Shunt}$ or Z_{Shunt} is larger than Z_P or R_P past the frequency where Z_{Shunt} equals R_P, which happens around unity-gain frequency f_{0dB}, when loop gain LG is 1:

$$Z_{Shunt} = \frac{1}{\beta_{FB} A_{EA} G_P} \approx \left.\frac{R_P}{LG}\right|_{f_{0dB}} \equiv Z_P\big|_{f < f_P} = R_P \qquad (5.43)$$

where A_{EA} refers to the error amplifier's gain and f_p the frequency when C_P shunts R_P, the latter of which occurs at frequencies higher than f_{0dB}.

Shunt impedance Z_{Shunt} continues to increase past f_{0dB}, causing impedance $Z_O \,||\, Z_{Shunt}$ to also increase. When Z_{Shunt} surpasses Z_O, however, as depicted in Fig. 5.17*d*, $Z_O \,||\, Z_{Shunt}$ reverses its trend and decreases

with Z_O, as output, bypass, and load capacitors C_O, C_B, and C_L short to ground. Impedance Z_O's impact on A_{IN} and PSR^{-1} is at first insignificant, however, because its value continues to be larger than R_P:

$$A_{IN.MF} = \left.\frac{Z_O \,||\, Z_{Shunt}}{Z_P + Z_O \,||\, Z_{Shunt}}\right|_{f>f_{Shunt}} \approx \frac{Z_O}{R_P + Z_O} \approx \frac{Z_O}{Z_O} = 1 \tag{5.44}$$

where f_{Shunt} is the frequency where Z_{Shunt} equals Z_O or $(f_{0dB}p_O)^{1/2}$:

$$Z_{Shunt} = \frac{1}{\beta_{FB}A_{EA}G_P} = \left.\frac{1 + sC_AR_A}{\beta_{FB}(G_AR_A)G_P}\right|_{f>>p_A} \approx \frac{sC_A}{\beta_{FB}G_AG_P}$$

$$= \left.\frac{sR_P}{2\pi f_{0dB}}\right|_{f_{Shunt}=\frac{1}{2\pi}\sqrt{\frac{2\pi f_{0dB}}{(C_O+C_B+C_L)R_P}}\approx\sqrt{f_{0dB}p_O}} \equiv Z_O\big|_{f>f_{0dB}} \approx \frac{1}{s(C_O + C_B + C_L)} \tag{5.45}$$

where gain-bandwidth product GBW, which approximately equals f_{0dB}, is approximately $\beta_{FB}G_AG_PR_P/2\pi C_A$. Note f_{Shunt} is between f_{0dB} and output pole p_O.

Supply injection again drops when $Z_{Shunt} \,||\, Z_O$ (or now just Z_O) decreases below resistance R_P (as illustrated in Fig. 5.17*e*), as R_P alone now sets the total series resistance between v_{IN} and ground—R_P + $(Z_O \,||\, Z_{Shunt})$ approximately reduces to R_P. The frequency where this shift in impedance dominance occurs roughly coincides with output pole p_O:

$$(Z_{Shunt} \,||\, Z_O)\big|_{f>f_{Shunt}} \approx Z_O\big|_{f>f_{Shunt}}$$

$$\approx \left.\frac{1}{s(C_O + C_B + C_L)}\right|_{p_O\approx\frac{1}{2\pi(C_O+C_B+C_L)R_P}} \equiv R_P \tag{5.46}$$

At some higher frequency f_P, however, feed-forward impedance C_P shunts R_P (as seen in Fig. 5.17*f*) and limits the extent to which A_{IN} or PSR^{-1} improves to $C_P/(C_O + C_B + C_L)$:

$$A_{IN.HF} = \left.\frac{Z_O \,||\, Z_{Shunt}}{Z_P + Z_O \,||\, Z_{Shunt}}\right|_{f>f_{Shunt}} \approx \left.\frac{Z_O}{Z_P + Z_O}\right|_{f>p_O} \approx \left.\frac{Z_O}{Z_P}\right|_{f>f_P}$$

$$= \frac{\frac{1}{s(C_O + C_B + C_L)}}{\frac{1}{sC_P}} = \frac{C_P}{C_O + C_B + C_L} \tag{5.47}$$

where $A_{IN.HF}$ refers to A_{IN} at high frequencies. The next event to occur as frequencies increase is C_O shorts with respect to its R_{ESR} (as shown in Fig. 5.17*g*) at z_{ESR},

$$\left.\frac{1}{sC_O}\right|_{z_{ESR}=\frac{1}{2\pi C_O R_{RSR}}} \equiv R_{ESR} \tag{5.48}$$

Impedance Z_O therefore reduces to R_{ESR}, increasing A_{IN} (and degrading PSR) at 20 dB per decade of frequency:

$$A_{IN.VHF} = \left.\frac{Z_O \,||\, Z_{Shunt}}{Z_O + Z_O \,||\, Z_{Shunt}}\right|_{f>p_p} \approx \left.\frac{Z_O}{Z_P}\right|_{f>z_{ESR}} = \frac{R_{ESR}}{\frac{1}{sC_P}} = sC_P R_{ESR} \tag{5.49}$$

where $A_{IN.VHF}$ is A_{IN} at very high frequencies. At even higher frequencies, C_B and C_L short R_{ESR} (as shown in Fig. 5.17*h*) and A_{IN} again flattens (i.e., improves). These latter effects, however, typically occur at considerably higher frequencies because bypass capacitors C_B are not normally present in internally compensated regulators and load capacitors C_L can be substantially low.

Since output capacitor C_O in internally compensated regulators should not exceed a specified maximum value to ensure p_O resides at sufficiently high frequencies and therefore guarantee stable conditions, application engineers normally employ high-frequency (i.e., low-ESR or equivalently low time-constant) capacitors for C_O, which is why z_{ESR} typically exceeds f_P. However, when R_{ESR} is relatively large and parasitic capacitance C_P is low, z_{ESR} may fall below f_P, in which case the high-frequency supply-gain limit changes to $A'_{IN.HF}$ or R_{ESR}/R_P:

$$A'_{IN.HF} = \left.\frac{Z_O \,||\, Z_{Shunt}}{Z_P + Z_O \,||\, Z_{Shunt}}\right|_{f>f_{Shunt}} \approx \left.\frac{Z_O}{Z_P + Z_O}\right|_{f>p_O} \approx \left.\frac{Z_O}{Z_P}\right|_{f>z_{ESR}} = \frac{R_{ESR}}{R_P} \tag{5.50}$$

The onset of f_P, when C_P shunts R_P, once again increases A_{IN} at 20 dB per decade of frequency, just as it did when z_{ESR} resided above f_P (in $A_{IN.VHF}$).

In summary, as in externally compensated circuits, a v_{OUT}-sensed negative-feedback loop presents a ripple-shunting impedance Z_{Shunt} at v_{OUT} with a value that is inversely proportional to feedback transconductance G_{FB} (i.e., Z_{Shunt} is $1/\beta_{FB}A_{EA}G_P$) and whose effects are prevalent at low-to-moderate frequencies, when Z_O is higher than Z_{Shunt}. In the internally compensated case, however, error-amplifier bandwidth p_A is the dominant low-frequency pole of the feedback loop, asserting its degrading effects on PSR performance first, at considerably lower

frequencies, until A_{IN} peaks and flattens near unity-gain frequency f_{0dB} (as described in $A_{IN.MF}$). Output, bypass, and load capacitors C_O, C_B, and C_L then dampen this peaking effect (i.e., improve PSR), after which feed-forward capacitor C_P and R_{ESR} (as C_O shorts) increase the supply ripple in v_{OUT}. This general behavior indicates the following:

- High error-amplifier bandwidth p_A extends low-frequency supply gain $A_{IN.LF}$ to higher frequencies and therefore improves PSR at low-to-moderate frequencies.
- A high unity-gain frequency f_{0dB} implies the first and worse A_{IN} peak is at higher frequencies (as described in $A_{IN.MF}$), in other words, improves PSR.
- High C_O, C_B, and C_L values dampen the first and worst peaking effect in $A_{IN.MF}$, also improving PSR.
- Low C_P values inject less high-frequency supply ripple to v_{OUT}, decreasing $A_{IN.HF}$ and again improving PSR.
- High R_{ESR} values increase the supply-noise content in v_{OUT}, increasing $A_{IN.VHF}$ and degrading PSR.

Miller Effect in Internally Compensated Regulators

As derived directly in Chap. 3 and later discussed in Chap. 4, the Miller effect is a manifestation of negative feedback (because a capacitor connected across an inverting amplifier establishes a negative-feedback loop). Ultimately, its effect on small-signal voltage gain is to decrease it past Miller pole p_M, which the well-known Miller theorem perceives directly as an increase in input capacitance and indirectly (because it does not explicitly discuss it within this context) as a decrease in output impedance. From the perspective of the output, the shunt-feedback effect of the Miller capacitor is to decrease output impedance Z_O from its low-frequency point, when the capacitor is an open circuit, to its high-frequency value, when the capacitor is a short circuit. Viewing the Miller effect in this fashion also explains how the small-signal gain decreases past p_M, when the output impedance starts to decrease with increasing frequencies.

Consider, for example, a Miller capacitor C_M connected across common-source transistor amplifier M_P and driven from a circuit whose source resistance is R_S, as illustrated in Fig. 5.19. Output impedance Z_O is the parallel combination of r_{dsP}, the resistance into transconductor g_{mP} or equivalently Z_{gmP}, and the series impedance across R_S and C_M:

$$Z_O = r_{dsP} \parallel Z_{gmP} \parallel \left(R_S + \frac{1}{sC_M} \right) \tag{5.51}$$

Noting C_M ultimately diode-connects M_P, R_S, and C_M voltage-divide the small signal present at M_P's gate and therefore degenerate g_m

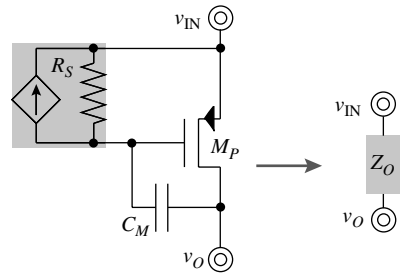


FIGURE 5.19
Miller capacitor C_M across inverting common-source PMOSFET M_P.

(at low-to-moderate frequencies) because only a fraction of v_O appears across M_P's gate-source terminals (i.e., v_{gsP} is a fraction of v_O). As a result, g_m in the $1/g_m$ resistance that M_P would have otherwise presented with a short circuit across its drain-gate terminals degenerates:

$$Z_{gmP} = \frac{\left(R_S + \frac{1}{sC_M}\right)}{g_{mP}R_S} \tag{5.52}$$

In other words, because C_M is an open circuit at low frequencies and the loop is open, Z_{gmP} approaches infinity and Z_O reduces to r_{dsP}. As frequency increases, the impedance across C_M decreases (although still considerably larger than R_S) and so does Z_{gmP}, except its impact on Z_O is negligible until Z_{gmP} decreases below r_{dsP}:

$$Z_{gmP} = \left.\frac{\left(R_S + \frac{1}{sC_M}\right)}{g_{mP}R_S}\right|_{\text{Low.Freq.}} \approx \left.\frac{\left(\frac{1}{sC_M}\right)}{g_{mP}R_S}\right|_{p_M \approx \frac{1}{2\pi r_{dsP} g_{mP} R_S C_M}} \equiv r_{dsP} \tag{5.53}$$

beyond which frequency Z_O decreases with Z_{gmP}, past Miller pole p_M at $1/2\pi r_{dsP} g_{mP} R_S C_M$, as also derived (by different means) in Chaps. 3 and 4. Past p_M, Z_O continues to decrease until C_M becomes a short circuit with respect to R_S (when C_M's impedance decreases below R_S):

$$\left.\frac{1}{sC_M}\right|_{f_{gm} \approx \frac{1}{2\pi R_S C_M}} \equiv R_S \tag{5.54}$$

beyond which point Z_O reduces to $1/g_{mP}$.

Within the context of Miller-compensated low-dropout (LDO) regulators, where capacitor C_M is across a power p-type transistor, as shown in Fig. 5.19, understanding the effect of C_M on the output

impedance is useful because the PSR analysis presented in this chapter reduces the output to a voltage-divider network from v_{IN} to v_{OUT}. With this in mind, as before, since the loop gain is high below unity-gain frequency f_{0dB}, shunting impedance Z_{Shunt} is substantially low and the equivalent PSR circuit reduces to the series combination of R_P and Z_{Shunt}, as shown in Fig. 5.17*a*, except R_P should really be impedance Z_P because R_P is no longer frequency independent (since C_M now changes it). Considering Z_{Shunt} is roughly the ratio of open-loop impedance Z_P (from Z_O in Fig. 5.19) and loop gain $A_{V.OL}\beta_{FB}$ and recalling both small-signal gain $A_{V.OL}$ and Z_P decrease with frequency past Miller pole p_M, Z_{Shunt} in Miller-compensated LDO regulators does not increase past p_M. Interestingly, C_M's overall impact on supply gain A_{IN} (and PSR) at low-to-moderate frequencies is unchanged with respect to the internally compensated case discussed earlier because A_{IN} is roughly the ratio of Z_{Shunt} and Z_P, and although the former remains unchanged past p_M, the latter does not, decreasing past p_M and consequently increasing A_{IN} past p_M. Note that past f_{gm}, when Z_P reaches $1/g_{mP}$, Z_P stops decreasing with frequency and Z_{Shunt} starts increasing (as $A_{V.OL}$ continues to decrease), but because A_{IN} is approximately Z_{Shunt}/Z_P during this frequency range, A_{IN} continues to increase (now with Z_{Shunt}) and f_{gm} has no impact on PSR.

In more general and perhaps more intuitive terms, Miller compensation in LDO regulators decreases the impedance from the output to the positive supply (i.e., decreases Z_T), a condition that is unfavorable for good PSR performance. The fact is that a Miller capacitor across an inverting amplifier establishes a shunt negative-feedback loop whose shunting effects to the input supply appear more and more as frequency increases, as Miller capacitor C_M shorts. So while the impedance to the supply (i.e., Z_T) is high at lower frequencies (because C_M is an open circuit), Z_T decreases with increasing frequencies, as shunt feedback asserts its influence, in the process increasing supply gain A_{IN} and degrading PSR.

5.2.4 Conclusions

Perhaps the most important message in this section is how to convert a circuit into its voltage-divider equivalent for power-supply rejection (PSR) analysis. The pass device, to start, feeds v_{IN} signals to v_{OUT} through R_P and C_P, or equivalently (when combined), through Z_P. Filter-load network Z_O, on the other hand, which consists of $R_{FB1} + R_{FB2}$, R_L, C_L, C_O, and C_B, shunts v_{IN} noise in v_{OUT} to ground. Shunt feedback around the error amplifier then helps Z_O shunt signals to ground by presenting loop-gain reduced impedance Z_{Shunt} to ground at v_{OUT}. Additionally, the architecture of error amplifier with respect to the pass device can introduce feed-through signals to v_{OUT} from v_{IN}, but loading the amplifier with ground-referenced mirrors when driving n-type pass devices (or supply-referenced mirrors in the case of p-type power transistors) circumvents many of these ill-fated effects.

In the end, the voltage-divider model that results is comparatively simple and capable of accurately predicting the PSR performance of the regulator across the entire frequency range, barring the unforeseen effects of any peculiarities in the actual transistor-level circuit.

Considering the voltage-divider network that results, the key to good PSR performance (and low supply gain A_{IN}) is high supply impedance and low ground impedance, the latter of which translates to high feedback-derived transconductance G_{FB} (i.e., $\beta_{FB}A_{EA}G_P$) and high output, bypass, and load capacitances. Because the dominant low-frequency pole of an externally compensated regulator is at v_{OUT}, error-amplifier bandwidth p_A is at moderate-to-high frequencies, extending the frequency range for which Z_{Shunt} or G_{FB}^{-1} is low. The internally compensated counterpart, on the other hand, suffers from considerably higher supply ripple injection because its dominant low-frequency pole is p_A, which must consequently reside at considerably lower frequencies. Parasitic equivalent series resistance (ESR) R_{ESR}, irrespective of compensation strategy, increases A_{IN} and consequently degrades PSR performance, but its effects are typically more pronounced in externally compensated circuits because large high-ESR capacitors are substantially cheaper than their high-frequency counterparts are.

Good PSR is intrinsic at low-to-moderate frequencies. On one end, as an example, the regulator may derive energy and power directly from a highly variable and unregulated input supply, as with 2.7–4.2 V Li-Ion batteries, so good line-regulation performance (LNR) and, by translation, low-frequency supply gain $A_{IN.LF}$ (i.e., LNR is equivalent to $A_{IN.LF}$) are critical. On the other hand, an off-line 20 kHz-to-5 MHz switching source may supply power to the linear regulator, which means good (i.e., high) moderate-frequency PSR is important. Unfortunately, the unity-gain frequency of most linear regulators, near the point where A_{IN} typically peaks and PSR is worst, is relatively low at 20 kHz to 1 MHz, which means the linear regulator is unable to suppress the switching noise injected by the off-line source. This limitation in bandwidth results because (1) R_{ESR} and R_P's respective expansive ranges (in response to process, temperature, and load variations) shift ESR zero z_{ESR}, bypass pole p_B, and output pole p_O considerably and (2) the low quiescent-current demands associated with battery-powered devices place severe restraints on the extent to which error-amplifier bandwidth p_A can reach. Nevertheless, because p_A is at higher frequencies, externally compensated circuits enjoy better moderate-frequency PSR performance than their internally compensated counterparts. The drawback, of course, is large off-chip capacitors often prohibit system-on-chip (SoC) and/or system-in-package (SiP) integration.

The values and shape of PSR response across frequency for externally and internally compensated circuits may differ but their underlying themes do not. Generally, increasing the impedance from v_{IN} to

v_{OUT} (i.e., increasing Z_P) and decreasing the impedance from v_{OUT} to ground (i.e., decreasing Z_O and Z_{Shunt}) reduces supply gain and yields better PSR performance. High low-frequency loop gain LG_{LF} and high error-amplifier bandwidth p_A, for instance, results in good low-frequency PSR values. Low R_{ESR}, high C_B, high R_P, and low C_P values produce similar effects at moderate-to-high frequencies. The difficulty is that higher loop gain and bandwidth require more battery drain current and low-impedance passives demand costlier process technologies and system solutions.

Note the analysis presented assumed the error amplifier only had one signal-shunting pole (i.e., p_A), which is reasonable, given the stability constraints of the system, but not entirely true. Parasitic poles in the amplifier may affect the PSR response of the circuit at moderate frequencies because they could change the characteristics of shunt impedance Z_{Shunt} and the common-mode nature of the amplifier's output (i.e., v_A) with respect to the pass device. At moderate frequencies, for instance, when Z_{Shunt} is still smaller than or equal to output impedance Z_O and the parallel combination of Z_{Shunt} and Z_O reduces to Z_{Shunt}, changes in Z_{Shunt} resulting from the parasitic poles in the amplifier have a direct impact on PSR. These effects are typically less significant because the additional poles reside above the regulator's unity-gain frequency (for stability purposes), around and beyond the frequency where Z_{Shunt} surpasses Z_O and PSR no longer depends on Z_{Shunt}. These same poles, however, also change the common-mode nature of error-amplifier output v_A, the result of which is a growing feed-through signal through the pass device and more noise in v_{out} at or above f_{0dB}.

5.3 External versus Internal Compensation

The overriding advantages of externally compensated regulators are high output capacitance C_O and high error-amplifier bandwidth p_A. High C_O values, on one hand, increase the regulator's ability to suppress fast and high-power load dumps and high p_A frequencies, on the other, extend good low-frequency PSR performance to higher frequencies. Maintaining p_A at high frequencies to guarantee stable operating conditions, however, is challenging, especially when considering the error amplifier must drive the parasitic capacitance the necessarily large power pass device presents. Internally compensated circuits reap the integration benefits of lower output capacitances, more easily conforming to the total on-chip integration demands of state-of-the-art portable, battery-powered solutions. The costs of on-chip integration, however, are lower output power and more stringent stability constraints, as output pole p_O has a tendency to migrate to lower frequencies during light loading conditions. Table 5.1 presents a summary of these comparative conclusions.

Low quiescent currents, when considering extended battery-life operation, necessarily pull parasitic poles to lower frequencies, limiting

	Externally Compensated	Internally Compensated
Dominant Pole	Output pole p_o	Error-amplifier pole p_a
C_o Limit	Greater than $C_{Specified}$	Less than $C_{Specified}$
Load-Dump Response	Better (i.e., lower Δv_{OUT})	Worse (i.e., larger Δv_{OUT})
Worst-Case Stability Conditions	↑ **f_{0dB}**: ↑ I_L, ↓ C_O, ↑ R_{ESR}, ↓ C_B' **2 poles, no z_{ESR}**: ↓ I_L, ↑ C_O, ↓ R_{ESR}, ↑ C_B'	↑ **f_{0dB}**: ↑ I_L, ↓ C_O, ↑ R_{ESR}, ↓ C_B' **2 poles, no z_{ESR}**: ↓ I_L, ↑ C_O, ↓ R_{ESR}, ↑ C_B'
PSR (or A_{IN}^{-1})	Better (i.e., higher p_A reduces $A_{IN.LF}$)	Worse (i.e., lower p_A increases $A_{IN.LF}$)
Integration	Off-chip or in-package C_O	On-chip or in-package C_O
Application	**Higher power (i.e., larger load dumps)**	**Lower power (i.e., smaller load dumps)**

TABLE 5.1 Externally versus Internally Compensated Regulators

the extent to which f_{0dB} should increase to maintain stability. To aggravate matters, the presence of z_{ESR} and p_B extends f_{0dB} and decreases the number of decades the loop gain can drop at 20 dB per decade from its low-frequency value, as seen in Fig. 5.3*b*, which results in relatively low loop gains when compared to standard op amps, and by translation, limited load- and line-regulation performance. Therefore, generally speaking, irrespective of the compensation strategy, increasing both loop gain and bandwidth under low quiescent-current restrictions compromises stability, which explains why loop gains normally fall below 50–60 dB and nominal unity-gain frequencies below 0.5–1 MHz.

5.4 Summary

The variations of load current I_L and equivalent series resistance (ESR) R_{ESR} are probably the two most significant obstacles to overcome when stabilizing a linear regulator. In a battery-operated environment, for instance, the former varies by up to five decades, from maybe 1 μA during light loading conditions to 100 mA when fully loaded. As a result, the location of output pole p_O also shifts by up to five decades, because p_O is inversely proportional to pass resistance R_P and therefore directly proportional to I_L. To complicate matters, although not to the same degree, R_{ESR} varies considerably across process and temperature (e.g., 1–500 mΩ). The consequence of this R_{ESR} variation is that the location of the ESR zero and bypass pole pair that results (i.e., z_{ESR} and p_B) can fall anywhere below or above unity-gain

frequency f_{0dB}, possibly extending f_{0dB} by one to two decades, depending on the ratio of p_B to z_{ESR} (i.e., ratio of C_O to $C_B + C_L + C_P$). The total possible variation for f_{0dB} can therefore be on the order of five to seven decades.

Load-dependent changes in low-frequency loop gain LG_{LF} sometimes offset the effects of a variable p_O, but only almost completely in internally compensated regulators. In Miller-compensated PMOS regulators, for example, the load-dependent component in dominant low-frequency pole p_A (i.e., $p_A \propto 1/C_{Miller} \propto 1/G_P R_P \propto I_L/I_L^{1/2}$ or $1/I_L^{1/2}$) complements that of low-frequency gain LG_{LF} (i.e., $LG_{LF} \propto G_P R_P \propto I_L^{1/2}/I_L$ or $I_L^{1/2}$). Internally compensated n-type regulators also yield a constant gain-bandwidth product because the gain across the pass device is close to unity and almost independent of I_L, as is dominant low-frequency pole p_A. Nevertheless, variations in p_O, if allowed to fall below f_{0dB}, can shift f_{0dB} by up to one decade. In externally compensated PMOS regulators, on the other hand, the gain across the pass device (i.e., $G_P R_P$) only partially offsets linear changes in p_O (where $p_O \propto I_L$) because $G_P R_P$ is inversely proportional to $I_L^{1/2}$, decreasing the overall frequency variance of f_{0dB} with respect to I_L by about half the decades (e.g., a four-decade change in I_L induces a two-decade shift in f_{0dB}). In PNP regulators, however, f_{0dB} shifts by the same number of decades as I_L because the gain across the pass device is independent of I_L, not in the least offsetting the effects of I_L on p_O.

The difficulties compound when considering high output, bypass, and load capacitances C_O, C_B, and C_L and high loop gain are desirable. High output filter capacitances, on one hand, attenuate transient voltage excursions in v_{OUT} and high loop gains, on the other hand, reduce the adverse effects dc and ac changes in load current I_L and supply voltage v_{IN} exert on v_{OUT} (through load regulation and transient response in the case of I_L and line regulation and power-supply rejection with v_{IN}). Unfortunately, high output filter capacitances pull p_O to lower frequencies and high error-amplifier gains normally require high-resistance nodes (i.e., low-frequency pole p_A), the combination of which may be a system with two low-frequency poles and no zeros to salvage phase (as z_{ESR} can be at high frequencies). Pushing p_A to higher frequencies is not trivial because the parasitic capacitance the power pass device presents to the error amplifier is substantially large, on the order of 50–100 pF.

Loop gain and power-supply rejection (PSR) are inextricably related. High loop gains, for instance, yield low supply gains that result in good line-regulation performance and high power-supply rejection at low-to-moderate frequencies. Most notably, error-amplifier bandwidth p_A incurs the first and probably most significant debilitating effect on supply rejection because p_A mitigates the shunting effects of the negative-feedback loop. High C_O, C_B, and C_L values ultimately dampen the supply-injection peak that results from a low-to-moderate p_A. High R_{ESR} (which results in a low z_{ESR}) and high

feed-through capacitance in the power pass device (i.e., high C_p), on the other hand, increase the supply-noise content in v_{OUT} at moderate-to-high frequencies, but high bypass capacitances (i.e., high C_B) offset these limits for the better.

Generally, given the voltage-divider nature of the equivalent supply-rejection circuit, increasing the supply impedance between v_{IN} to v_{OUT} (e.g., increasing R_p and $1/sC_p$) and decreasing the ground impedance from v_{OUT} to ground (i.e., decreasing R_{FB1}, R_{FB2}, R_L, R_{ESR}, $1/sC_O$, $1/sC_B$, $1/sC_L$, and Z_{Shunt} or $1/\beta_{FB}A_{EA}G_p$) decrease supply gain and therefore increase supply rejection. The fact is the pass device in a linear regulator circuit *is* the fundamental feed-through component of v_{IN} noise, assuming, of course, the circuit does not unnecessarily introduce other feed-through paths to v_{OUT}. Complying with the common-mode requirements of the pass device, for example, by loading the differential pair of the error amplifier with an appropriately referenced mirror avoids feeding additional supply noise to v_{OUT}.

Ultimately, understanding ac constraints is important to ensure transistor- and layout-level design choices do not otherwise compromise the stability and/or contrast the basic performance objectives of the regulator. From here, the next step in the analog design process is to develop a transistor-level circuit that achieves the functionality described in Chap. 1 and comprehends the ac requirements just presented. The following chapter therefore examines how to employ the semiconductor transistors, resistors, capacitors, and basic building blocks introduced in Chaps. 2 and 3 to design and build an integrated circuit that (1) regulates v_{OUT} against, among other factors, dc and ac variations in supply and load; (2) meets the negative-feedback requirements discussed in Chap. 4, and (3) achieves the compensation and supply-rejection goals identified here in this chapter.

CHAPTER 6
IC Design

The purpose of a series linear regulator is to produce and maintain an independent, stable, noise-free, and predictable output voltage v_O. Its output must therefore be low impedance and derived from an accurate dc reference voltage V_{REF}. As such, referring to the compartmentalized regulator shown in Fig. 6.1, the general design approach is to employ negative feedback and exploit the regulation benefits gained from shunt sampling v_O with feedback resistors R_{FB1} and R_{FB2} and series mixing output-derived feedback signal v_{FB} and V_{REF} with differential error amplifier A_{EA}.

The design process of a typical linear regulator, as with most other electronic systems, begins with its target specifications, and more specifically, with its load and the power transistor that drives it. To start, the targeted dc output voltage prescribes the gain ratio required from output feedback resistors R_{FB1} and R_{FB2} with respect to V_{REF}, given the series-mixed feedback loop approximately equates v_{FB} to V_{REF}. DC conditions like input voltage V_{IN}, load or output current I_O, and dropout voltage V_{DO} then set the basic operating requirements of the series pass device (i.e., power switch S_O). After that, load and line ac accuracies in the presence of nonidealities in output capacitor C_O (like finite capacitance and nonzero equivalent series resistance R_{ESR}) determine how quickly power transistor S_O must react to transient variations in v_{IN} and i_O.

Because the operating specifications often require S_O to conduct high currents, S_O is large and its input control terminal substantially capacitive. Low-impedance buffer circuit A_{BUF} must therefore drive S_O with sufficient current and voltage reach (i.e., swing) to charge and discharge the large parasitic capacitance present within an acceptably low timeframe. To this end, A_{BUF} demands circuit-specific attributes of error amplifier A_{EA}, other than the ones dc and ac accuracy parameters like input-referred offset V_{OS}, load and line regulation (LDR and LNR), signal-to-noise ratio (SNR), and power-supply rejection (PSR) already impose.

With the design of the regulator in mind, the underlying objective of this chapter is to connect the system-level specifications introduced in Chap. 1 to practical integrated-circuit (IC) solutions. The initial

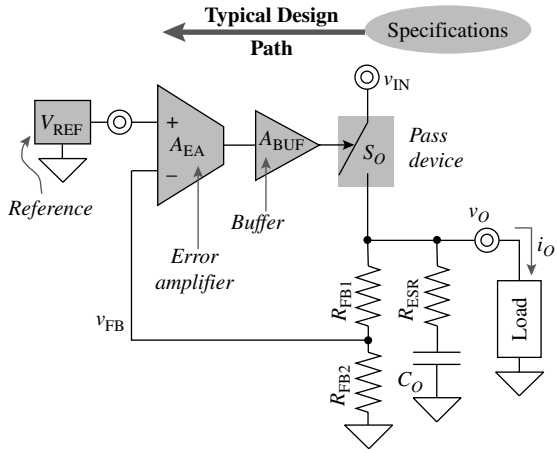


FIGURE 6.1 Compartmentalized linear-regulator circuit.

approach is to compartmentalize the design into its constituent components: (1) series power pass device S_O, (2) buffer A_{BUF}, and (3) error amplifier A_{EA}, and design them according to the specifications of the system and their interdependent parameters. Because the goal is to sustain a specified load, the design process starts at the output with S_O and ends with A_{EA}, as the driving demands ripple back through the system. Each section in the chapter discusses the various design considerations and associated physical-layout implications that practical loads impose on the IC with emphasis on efficiency and regulation performance.

6.1 Series Power Pass Device

The design goals of power switch S_O are to drive and accommodate a wide-ranging steady-state load (i.e., dc current I_O) with sufficient power efficiency to sustain acceptable single-charge battery lifetimes. Increasing efficiency η_P amounts to reducing the power lost across and because of S_O, which means avoiding ground current I_{GND} and decreasing supply voltage V_{IN} so that V_{IN} lies just above V_O, especially when supplying peak load $I_{O(max)}$:

$$\eta_P \equiv \frac{P_O}{P_{IN}} = \frac{P_O}{P_S + P_O} = \frac{P_O}{(V_{IN} - V_O)I_O + V_{IN}I_{GND} + P_O} \leq \frac{P_O}{V_{DO}I_O + V_{IN}I_{GND} + P_O} \tag{6.1}$$

where P_O is output power $V_O I_O$, P_S the power lost in and because of the switch, and V_{DO} the dropout voltage, which refers to the minimum voltage across S_O (when the loop gain falls to zero and S_O enters

the ohmic region, in other words, when the regulator is in dropout). To complicate matters, while driving high power with high efficiency, the power device cannot afford to significantly load its driving buffer with capacitance because doing so would otherwise slow the circuit, the effects of which compromise stability and degrade ac-accuracy performance in response to fast load dumps. As a result, S_O must operate under a low supply voltage, produce a low dropout voltage, and use little to no ground current, and when considering the load, be only large enough to sustain $I_{O(\max)}$ under worst-case process, temperature, and driving conditions.

6.1.1 Alternatives

Power pass device S_O is simply a transistor, since a negative-feedback loop need only modulate the transistor's input control terminal to control the resistance (and by translation, the conductance) across V_{IN} and v_O. Although a junction field-effect transistor (JFET), by definition, accomplishes this task, increasing the JFET's resistance so that little to no current flows (i.e., shut the transistor off), which is especially important in battery-supplied systems during idle conditions, is difficult. Bipolar-junction transistors (BJTs) and enhancement metal-oxide-semiconductor field-effect transistors (MOSFETs), on the other hand, are normally off devices that only require standard dual-supply rails (e.g., V_{IN} and ground) to induce substantial current flow. Because positive (or negative) input voltages with respect to the current-sourcing (or current-sinking) terminal induce current flow in both n-type (or p-type) BJTs and MOSFETs, the following subsections split the discussion into n- and p-type switches, rather than BJTs and MOSFETs.

N-type Power Pass Devices

Both npn BJTs and n-type MOSFETs, as required by the positive-supply nature of a linear regulator (where V_O is above ground), have their respective current-sourcing terminals (i.e., emitters and sources) attached to v_O, as illustrated in Fig. 6.2. This configuration carries an intrinsic advantage in that v_O is already low impedance (without the

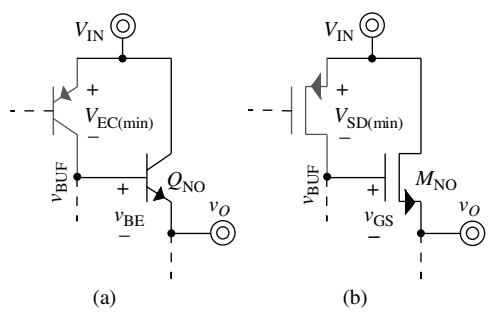


Figure 6.2 N-type (*a*) BJT and (*b*) MOSFET power-pass transistors.

help of the outer negative-feedback loop shown in Fig. 6.1), meaning power transistor Q_{NO} or M_{NO} responds quicker to quick changes in load (i.e., load dumps)—Q_{NO}'s or M_{NO}'s inherent output resistance is approximately equivalent to the reciprocal of its transconductance. The BJT version offers slightly better transient response than the MOSFET because the BJT's output resistance is even lower, as its collector current is exponential with respect to base-emitter voltage v_{BE} and the MOSFET drain current only follows the square of gate-source voltage v_{GS}.

The main disadvantage with n-type transistors is high-dropout (HDO) voltage. To be more specific, the minimum headroom required to induce $I_{O(max)}$ to flow between V_{IN} and v_O in an npn BJT (or NMOSFET) is the sum of its v_{BE} (or v_{GS}, which is the sum of threshold voltage V_T and saturation voltage $V_{DS(sat)}$) and the minimum emitter-collector voltage $V_{EC(min)}$ (or source-drain $V_{SD(sat)}$) of its driving transistor. In other words, V_{IN} must exceed v_O by 0.2–0.3 V above a diode voltage (or 0.4–0.8 V above V_T) to operate properly, which equates to dropout voltages V_{DO} on the order of 0.7–1.2 V (or 0.9–1.5 V). Nevertheless, the fast response time associated with these devices often outweighs the efficiency disadvantage that results from a larger V_{DO}. Note the BJT's base current flows to the output through the emitter so no current is lost as ground current I_{GND}.

P-type Power Pass Devices

In contrast to n-type devices, the current-sourcing terminals attached to v_O in p-type configurations are collectors and drains (Fig. 6.3), which means, unlike their n-type counterparts, they offer relatively high output-impedance characteristics (i.e., r_o for BJTs and r_{ds} for MOSFETs). The lower dropout voltage (and higher efficiency) associated with this setup, however, offsets the high-impedance disadvantage because V_{IN} need only be one $V_{EC(min)}$ or $V_{SD(sat)}$ above v_O to sustain $I_{O(max)}$ without collapsing power device Q_{PO} or M_{PO} into the ohmic region, all of which translates to a lower V_{IN} and dropout voltages of roughly 0.2–0.4 V. To fully source $I_{O(max)}$, though, V_{IN} must also exceed ground by at least one v_{EB} above $V_{CE(min)}$ or one v_{SG} above $V_{DS(sat)}$ (i.e., two $V_{DS(sat)}$'s higher than V_T), or approximately 0.7–1.5 V. Note that,

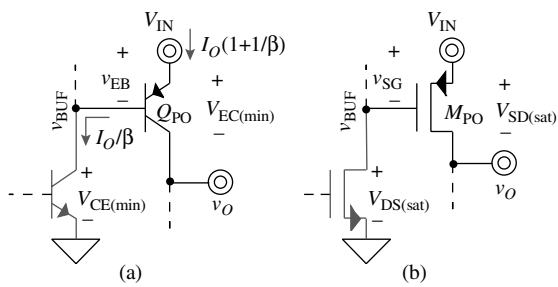


Figure 6.3 P-type (*a*) BJT and (*b*) MOSFET power pass transistors.

unlike the npn follower, a pnp transistor's base current flows to ground (not v_O), drawing and losing $I_{O(max)}/\beta$ from V_{IN} to ground as I_{GND}, which can be substantial under high temperatures and weak process-corner conditions (i.e., low β).

Comparative Evaluation

As with most designs, choosing the appropriate pass device is plagued with tradeoffs. The importance of high-efficiency performance in battery-powered microelectronics (to extend battery life) and high-power applications (to avoid the need for heat sinks) accentuate the value of low dropout voltages (i.e., V_{DO}) and low ground currents (i.e., I_{GND}). As a result, when considering the available choices, as summarized in Table 6.1, p-type transistors offer the lowest V_{DO} and MOS devices the lowest I_{GND} (because there is no dc current flow through a MOS gate), which explains why, when combining these considerations, PMOS pass devices are increasingly popular in the industry. Although not always true, low input-voltage requirements (i.e., $V_{IN(min)}$) appear hand in hand with low dropout voltages so similar conclusions apply when designing for low $V_{IN(min)}$. Nevertheless, the inherent low-impedance and therefore fast-response attributes of n-type power devices appeal to applications that value faster response times over efficiency performance.

BJTs, because of their exponential behavior, typically source higher $I_{O(max)}$ than MOSFETs under similar V_{IN} and silicon real-estate constraints. Optimized vertical pnp transistors, however, are seldom available and their lateral counterparts (which are available in standard single-well CMOS technologies) are usually less ideally suited to supply $I_{O(max)}$ (i.e., high currents) and respond quickly to changes in I_O (i.e., to load dumps). NMOS devices source relatively less current than their p-type counterparts do because V_O constrains their gate drives with respect to V_{IN} (e.g., $V_{IN} - V_O$). Again, the speed associated with NMOS followers partially offsets high V_{DO} and low I_O performance.

	N-type Power Pass Devices		P-type Power Pass Devices	
Parameter	**BJT**	**MOS**	**BJT**	**MOS**
$V_{IN(min)}$	$V_O + V_{DO}$	$V_O + V_{DO}$	$V_{EB} + V_{CE(min)}$ √	$V_{SG} + V_{DS(sat)}$
V_{DO}	$V_{EC(min)} + V_{BE}$	$V_{SD(sat)} + V_{GS}$	$V_{EC(min)}$ √	$V_{SD(sat)}$
I_{GND}	0 A √	0 A √	$I_{O(max)}/\beta$	0 A √
$I_{O(max)}$	Highest √	Low	High	Moderate
R_O	$1/g_{m(BJT)}$ √	$1/g_{m(MOS)}$	r_o	r_{ds}

Check marks "√" indicate preferred performance

TABLE 6.1 Series Power Pass Devices

Circuit Parameter	Specification	PMOS Process Parameter	Value
V_{IN}	0.9–1.6 V	$\lvert V_{TP}\rvert$	0.6 V ± 150 mV
V_O	1 V	K'_P	40 μA/V² ± 20%
I_O	≤ 50 mA	L	≥ 0.35 μm
V_{DO}	≤ 200 mV		
I_{GND}	0 A		

TABLE 6.2 Target Specifications and Process-Parameter Values for the Power Device in Example 6.1

Design Example 6.1 Use the process parameters outlined in Table 6.2 to design a power pass device with 200 mV of dropout and no ground current that delivers up to 50 mA to a load whose voltage must remain at 1 V when supplied from a 0.9–1.6-V NiMH battery.

1. For zero I_{GND} and a low V_{DO}, use a PMOSFET.

2. $$V_{DO} \le R_{DS(max)} I_{O(max)} \approx \frac{I_{O(max)}}{K'_{P(min)}\left(\frac{W}{L}\right)(V_{IN(min)} - |V_{TP(max)}|)} \le 100 \text{ mV}$$

$$\therefore \quad \frac{W}{L} \ge \frac{50 \text{ m}}{(32\ \mu)(200 \text{ m})(0.9 - 0.75)} = 52{,}083 \qquad \text{or} \qquad \frac{W}{L} \equiv \frac{18{,}300\ \mu\text{m}}{0.35\ \mu\text{m}}$$

3. $$V_{DS(sat)} \le \sqrt{\frac{2I_{O(max)}}{K'_{P(min)}\left(\frac{W}{L}\right)}} = \sqrt{\frac{2(50 \text{ m})}{(32\ \mu)\left(\frac{18{,}300}{0.35}\right)}} \approx 244 \text{ mV}$$

∴ The worst-case onset of regulator's low-gain mode (while still in the linear region, in other words, not in dropout) is when $V_{IN} \approx V_O + V_{DS(sat)} \approx 1.244\text{V}$.

6.1.2 Layout

Reliability and performance drive the physical layout considerations of a power transistor. High-power conditions, for example, if designed to reduce silicon real-estate overhead, stress devices to their safe-operating-area (SOA) limits, beyond which point the power switches begin to break down. As a result, the physical design of the device must account and avoid occurrences of electron migration, hot spots on the IC that lead to secondary breakdown events, substrate debiasing that leads to further debiasing and thermal runaway, and other adverse effects. Typical high-power considerations therefore include the current-density limits of current-carrying circuit paths (i.e., metal widths) and current-crowding effects within the device, the latter of

which ballasting resistors address by evenly distributing current (at the cost of additional ohmic power losses). However, given the low-dropout (i.e., high-efficiency) demands of portable, battery-powered applications, high-performance requirements often cause the size of the power pass device to exceed the SOA needs of the application.

With respect to dropout, the aim is to reduce parasitic resistances in the power-conducting path, as they not only dissipate power and increase dropout voltage but also generate heat. In practice, ICs introduce parasitic ground-current paths whose effects exacerbate under high-power conditions and in large surface-area devices. Inducing this ground current further decreases power efficiency and propagates noise through the common substrate, also increasing the propensity for latch-up. In attenuating these adverse effects, the general layout approach is to increase the impedance through these parasitic paths and, for the current that leaks through, provide low-impedance paths to the supplies, in other words, steer current away from the common substrate.

In light of the plethora of process technologies available and their possible mask-set compositions, there are many flavors of a particular transistor to consider. The preponderance of junction-isolated standard bipolar, vanilla CMOS, and basic biCMOS technologies, however, which have a common but reverse-biased silicon substrate, reduce layout approaches to a few general strategies: (1) vertical BJTs, (2) lateral BJTs, (3) substrate MOSFETs, and (4) welled MOSFETs. CMOS-based biCMOS (i.e., p-substrate, n-well CMOS with a p-base-diffusion layer) and standard p-substrate, n-epitaxy bipolar technologies, for instance, offer vertical npn and lateral pnp structures. Fully complementary bipolar processes offer vertical npn and pnp BJTs as well as lateral BJTs, although vertical structures normally outperform their lateral counterparts. In the same vein, n-well CMOS technologies offer substrate n-type and welled p-type transistors whereas their p-well complements have substrate p-type and welled n-type devices. Dual-well, p-substrate CMOS process technologies, on the other hand, offer both welled n- and p-type devices as well as substrate NMOS transistors. Given the prevalence of these four variants, the intent of the following subsections is to highlight the driving physical considerations associated with each basic layout strategy by citing and discussing common structures in view of their dropout performance and ground-current requirement.

Vertical BJT

Vertical npn BJTs, as shown in the profile view of Fig. 6.4*a*, are often available in standard bipolar and basic biCMOS technologies. The n-type minority carriers in such a device constitute power collector current i_C, flowing vertically from emitter v_O (at the silicon surface) to the intrinsic collector (i.e., n+ buried layer) and then sideways and up to extrinsic collector terminal V_{IN}. As such, to decrease the resistance

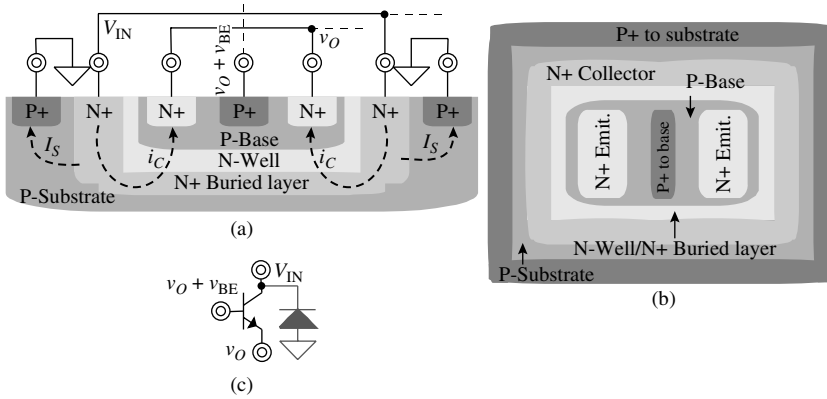


FIGURE 6.4 (*a*) Profile, (*b*) top views, and (*c*) equivalent schematic of a typical p-substrate vertical npn power pass device available in standard bipolar and basic biCMOS process technologies.

and consequential dropout voltage of the power-conducting path, the emitter contact is close to the collector so the physical distance (which translates to resistance) minority carriers traverse (i.e., i_C) from v_O to reach V_{IN} is minimal. A ring of deep n+ collector plugs and an expansive n+ buried layer in the conduction path further decrease this series resistance.

Referring to the physical top view of a vertical BJT illustrated in Fig. 6.4*b*, increasing the number of emitter strips increases the power rating of the device, that is, the amount of current it conducts when forward-biasing its base-emitter junction. Inserting thin base strips between the emitters, on the other hand, decreases the series base resistance of the BJT, which is important to increase the large-signal speed (i.e., bandwidth) of the transistor because completely shutting down or fully engaging the device requires a substantial (*and* uniform) reduction or increase in base-emitter voltage. The deep n+ plug and corresponding collector contact surround (or "ring") the entire device to reduce the negative effects of series collector resistance on dropout voltage and minimum collector-emitter voltage $V_{CE(min)}$.

The vertical BJT structure shown only presents one reverse-biased diode from the collector (i.e., from V_{IN}) to the substrate (Fig. 6.4*c*) that, other than conducting reverse-saturation current I_S, induces little additional substrate current. The p-base, n-collector, and p-substrate regions also constitute a parasitic vertical pnp transistor, except its effects are minimal because the effective emitter-base junction of the parasitic device (i.e., base-collector junction of the power npn BJT) is typically reverse biased. Nonetheless, given the high-power nature of the transistor, extending the n+ buried layer to the edge of the n-well region is ideal because overlapping the entire base (Fig. 6.4*b*) provides electrons with which parasitic minority-carrier holes can recombine

before reaching the substrate (effectively decreasing the effective β of the parasitic vertical pnp BJT). A highly doped p-type guard ring around the entire device (as shown in Fig. 6.4*a* and *b*) collects and prevents any I_S that leaks into the common substrate from propagating to the rest of the IC.

Lateral BJT

Power lateral pnp BJTs available in mainstream CMOS and basic biCMOS process technologies, as illustrated in Fig. 6.5*a*, conduct power current i_C laterally from emitter to collector terminals so the periphery and diffusion depth of their respective emitters set the effective current-conducting areas of the device (not the surface area of the emitters). As a result, to collect as much current as possible, the collector (i.e., v_O) usually rings the emitter (i.e., V_{IN}), as shown in the figure. Unfortunately, the p-type emitter and collector terminals, the n base, and the p substrate also constitute parasitic vertical pnp devices Q_{DO} and Q_{SUB} from v_O and V_{IN} to ground (Fig. 6.5*a* and *b*). Because the collector-base junction of the intrinsic device is reverse biased, though, the parasitic device attached to v_O (i.e., Q_{DO}) is normally off, unlike the one attached to V_{IN} (i.e., Q_{SUB}) whose emitter-base terminals are shared with those of the intrinsic BJT, which means Q_{SUB} conducts undesired substrate current I_{SUB} to ground.

Decreasing the surface area of a unit emitter to the minimum permissible point the process technology allows favors intrinsic lateral

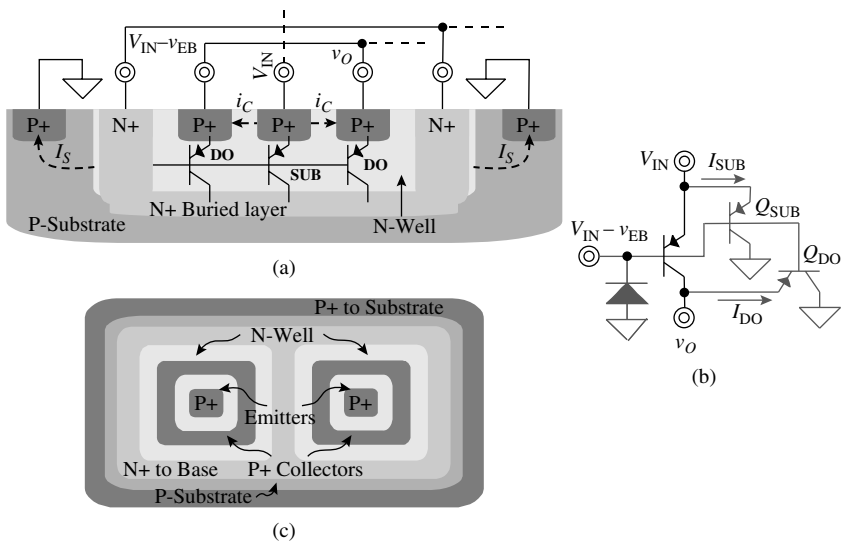


Figure 6.5 (*a*) Profile view, (*b*) equivalent schematic, and (c) top view of a typical p-substrate lateral pnp power pass device available in vanilla CMOS and basic biCMOS process technologies.

(i.e., peripheral) current i_C over parasitic vertical (i.e., surface area) substrate current I_{SUB}, in other words, increases the effective β of the lateral transistor, and paralleling numerous minimum emitter-dot pnp BJTs (Fig. 6.5*c*) increases its power reach. An n+ buried layer in the n-well that encloses the lateral BJT, as in the vertical npn case, further reduces I_{SUB} by supplying electrons with which the parasitic vertical minority hole carriers can recombine. Note that during dropout conditions, when the collector-base junction of the lateral transistor forward biases, the parasitic vertical device attached to v_O also forward biases and carries additional dropout current I_{DO} to ground as I_{SUB}. Although the n+ buried layer also helps decrease the gain of this parasitic BJT and its resulting substrate current, the buried layer actually steers the current back to the lateral BJT's base, ultimately losing it as ground current (i.e., quiescent power) through the buffer driver. In the end, lateral pnp BJTs are not ideally suited for a battery-powered linear power pass device because its ground-current losses are inherent and relatively substantial when compared against other devices.

The parasitic current paths may decrease efficiency but they also impede current crowding and the formation of hot spots on the IC, naturally increasing the ruggedness of the device. Ringing shallow and deep n+ contact base regions around each collector-emitter dot combination decreases the base resistance of the device and consequently improves its response time when driven to sustain large load-dump events. This ring is especially important in the absence of a buried layer, as in the case of vanilla CMOS and basic biCMOS technologies. P+ substrate contacts around the entire device collect and steer reverse-bias current I_S away from the rest of the IC.

Substrate MOSFET

Standard MOSFETs, as exemplified by the substrate NMOS device shown in Fig. 6.6*a*, conduct drain current i_D laterally so the source-drain separation (i.e., channel length) must be short and the conducting diffusion width (i.e., channel width) wide to present minimal series resistance to i_D and incur minimal additional dropout voltages. From a top-view perspective, the lateral length (when referring to Fig. 6.6*b*) of each source/drain finger (i.e., channel width) should therefore be long, but not extraordinarily so. An excessively long finger, by translation, increases the series gate resistance of the transistor and consequently degrades the large-signal response time of the device, as the resulting RC across the gate delays the drain-current response at the extreme ends of the finger strip.

Typical polysilicon-gate layouts resemble a fork or a tree structure with multiple but equally sized branches. Increasing the gate area increases the parasitic capacitance the transistor presents to the driving buffer, which decreases the response time of the regulator, so its base width (i.e., fork base or tree top) should not be excessive.

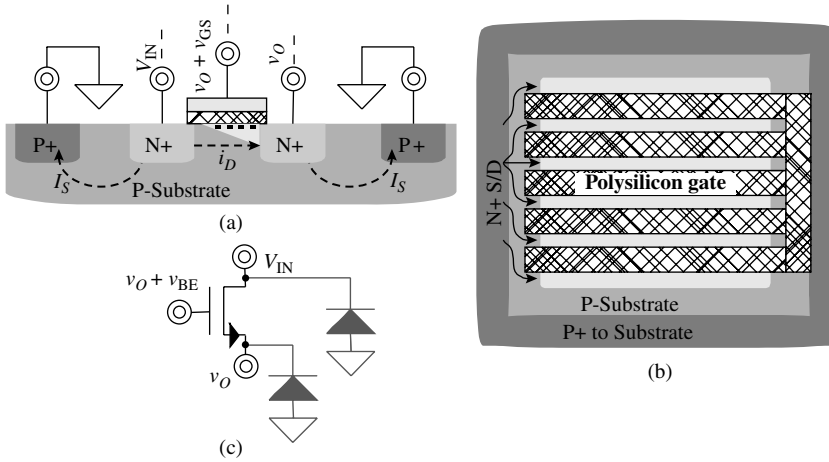


FIGURE 6.6 (*a*) Profile, (*b*) top views, and (*c*) equivalent schematic of a typical p-substrate NMOS transistor available in a standard CMOS technology.

The best means of increasing the power rating of the device is to maximize the length of each finger (e.g., 300–400 μm) to the point where series gate resistance just reaches its limit and increasing the number of source/drain fingers to match the power needs of the application. If numerous fingers are necessary, mirroring fingers on either side of the polysilicon base (about the gate vertebrae) minimizes the gate resistance of the entire structure.

The only parasitic devices present in this structure are reverse-biased diodes to substrate, as illustrated in the schematic of Fig. 6.6*c*. Other than conducting reverse-saturation current I_S, these diodes conduct little else. In spite of this, given the high-power nature of the device, a p+ substrate contact region normally rings the power transistor to steer all parasitic current I_S to ground directly, bypassing the common substrate.

Welled MOSFET

A welled MOSFET is also a lateral device (Fig. 6.7*a*) so its series resistance is lower if the source-diffusion distance (i.e., channel length) is short and the diffusion width (i.e., channel width) wide. The polysilicon gate should also mimic a fork or a tree when viewed from the top (Fig. 6.7*b*). The only difference between a substrate MOSFET and its welled counterpart is its back gate, being the substrate in the former and the well in the latter.

In spite of the seemingly benign difference, embedding a PMOS device in an n-well presents, as in the case of the lateral pnp BJT, parasitic vertical pnp transistors from source and drain diffusions through the well (as the base region) to the collecting substrate (which is at ground), as illustrated in Figs. 6.7*a* and *c*. Unlike the lateral pnp

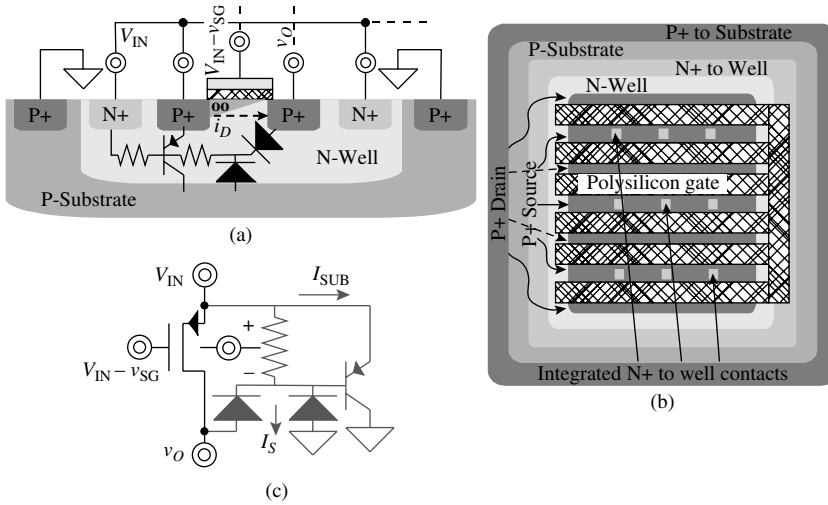


Figure 6.7 (*a*) Profile and (*b*) top views and (*c*) equivalent schematic of a typical p-substrate, n-well power PMOS transistor available in a standard CMOS technology.

device, neither parasitic BJT is intrinsically forward biased, and the one attached to v_O is reverse biased. Unfortunately, because the power transistor conducts substantially high currents and the parasitic device attached to V_{IN} is not really reverse biased (just short circuited via a resistive well), the well is prone to exhibit sufficient debiasing conditions (i.e., voltage drops) to induce the parasitic pnp transistor to conduct substrate current I_{SUB} to ground. Under lower currents, the debiasing voltage is not high enough to present a problem but subjecting a large device to high-power conditions may incur sufficient reverse-saturation current I_S through its reverse-biased diodes to substrate and to drain terminal v_O to debias the resistive well and forward bias the parasitic BJT, producing I_{SUB}.

As in the lateral BJT case, inserting a buried layer reduces the β of the parasitic vertical BJT, except a buried layer is not typically available in standard CMOS technologies. The next strategy is to reduce the series resistance between the extrinsic and intrinsic base terminals of the pnp transistor, that is, between V_{IN} and the well immediately around and below the p+ source diffusion. Conventionally, in low-power applications, butting the p+ source terminal against the n+ well contact region decreases this series resistance. Unfortunately, inserting an n+ well contact region next to every source finger in a high-power PMOS array demands considerable silicon area, which not only costs money but also presents higher overall capacitance to the driving buffer. Alternatively, integrating back-gate contacts into the source finger, as shown in Fig. 6.7*b* and later illustrated in Figs. 6.8 and 6.9,

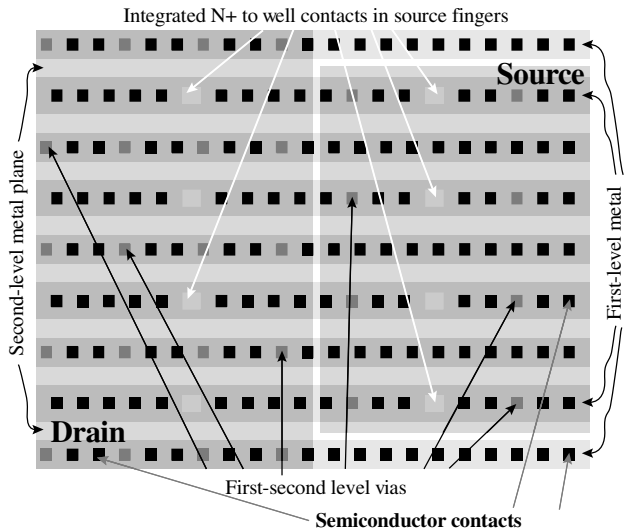


FIGURE 6.8 Top view of a vertical-metallization strategy for a welded power MOSFET from the semiconductor interface to second-level metal planes.

also decreases this resistance. Although doing so increases the finger width (i.e., lateral diffusion length) of the source slightly and prompts the layout tool to signal warnings, integrated back-gate contacts ultimately yield substantial real-estate savings at little to no cost in performance and reliability.

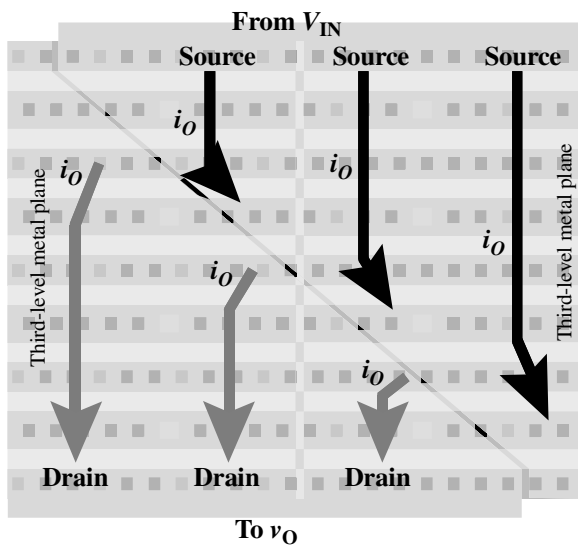


FIGURE 6.9 Top view lateral metallization strategy for a welded power MOSFET from second- to third-level metal planes and bond pads.

Metallization

Series metallic resistance R_M between power device S_O and the load and bond-wire resistance R_{BW} between the pin and bond pad increase the effective dropout voltage (i.e., V_{DO}) of a linear regulator:

$$V_{DO} = I_O R_{ON} = I_O(R_{SW} + R_M + R_{BW}) \tag{6.2}$$

where I_O is the output current and R_{SW} is the effective resistance of S_O when S_O is fully engaged to conduct current. Note R_M is the combined resistance from V_{IN}'s bond pad to S_O's input terminal and from S_O's output terminal to v_O's bond pad, just as R_{BW} refers to both V_{IN} and v_O's bond-wire resistances. The resistance attached to V_{IN} however, has more impact on V_{DO} in a p-type transistor than in an n-type counterpart because the critical dropout path in the former is across its emitter-collector or source-drain terminals, unlike the latter whose dropout path is through its base-emitter or gate-source terminals. Nevertheless, irrespective of the type, power devices are physically large so contacting their respective current-carrying terminals inevitably introduces series resistance in the power-conduction path. Given the power device introduces, by design, low R_{SW}, R_M, and R_{BW} may in some cases equal or overwhelm R_{SW} (like when the switch presents 50 mΩ, metallic links 25 mΩ, and bond wires 50 mΩ), which may ultimately imply the silicon area dedicated to the power transistor unnecessarily increases the cost of the device.

There are several strategies to decrease series metallization resistance R_M, and they all start and finish where the power device ends: at the silicon interface. The general V_{IN}-to-v_O approach is to (1) steer output current i_O laterally from input bond pad V_{IN} to a low-impedance plane (or planes) above power transistor S_O, (2) channel i_O down vertically to S_O's input terminal, (3) direct i_O up vertically from S_O's output terminal to another low-impedance plane (or planes), and (4) guide i_O laterally to output pad v_O. To decrease the vertical resistance, first-level metal covers the entire ohmic surface area of each emitter, collector, source, and drain areas. First-to-second level vias carry the current vertically to low-impedance second-level metal buses, as illustrated in the PMOS transistor example shown in Fig. 6.8.

Integrating the first-to-second level vias into the first-level metal fingers decreases the first-to-second metal interface resistance. Some process technologies now allow vias to sit directly above and overlap the semiconductor-metal contacts, which means integrating vias into the fingers incurs little to no tradeoffs with respect to semiconductor-to-metal resistance, as vias do not take the place of semiconductor contacts. Note the PMOS source fingers include integrated n+ back-gate (i.e., n-well) contacts to decrease n-well debiasing effects, as discussed earlier, which is why the width of the source fingers is slightly wider than the drain fingers' is. The ratio of source-diffusion to back-gate contacts is an engineering choice whose lower limit should not, in

practice, fall below roughly four source-diffusion contacts to each well contact to keep series resistance across the conduction path low.

In steering the current laterally to the bond pads, it is important to note a lateral metal track above the transistor "picks up" (or "drives down") parts of its current as it traverses across the IC. When sourcing current from the output terminal of a power transistor to output pad v_O, for instance, current increases as the metal plane carrying the current derives more of its current from underlying current-carrying semiconductor regions. If this lateral metal plane has a constant width, the current density throughout the plane is inconsistent, which leads to current crowding and inefficient use of the area available. As a result, increasing (or decreasing) the width of the sourcing (or sinking) plane as its current increases (or decreases), as shown in Fig. 6.9, is usually best. Depending on the ultimate location of the bond pads and the number of metal layers used, current may grow or fade in a number of different directions, giving the top-view layout geometry of a power device a multitriangular look, as triangles represent equal current-flux regions carrying increasing or decreasing currents. The metal traces across a high-power device seldom appear rectangular, unless current levels are so low that they incur negligibly low series voltage drops.

The power device should ideally sit close to the input and output bond pads of the regulator, like in the corner of a large IC, to minimize the lateral length (i.e., resistance) of the current-carrying top-level metal planes. If placing the device immediately next to its bond pads is impossible, multiple metal planes in parallel with the top-level plane, which usually has the lowest sheet resistivity, should carry the current to the pads. If more than three levels of metal are available, they should parallel the efforts of the other planes with the objective of decreasing series resistance (i.e., dropout voltage) and balancing (i.e., ballasting) current flow to maintain current densities low and consistent through the entire device. Similarly, when possible, using the pins in the lead frame that are closest to their respective bond pads and allocating multiple bond wires (and/or bond pads) in parallel for V_{IN} and v_O decrease bond-wire resistance R_{BW}.

There are several ways to include over-current (i.e., short-circuit) protection directly with series or parallel current sensors or indirectly through thermal-shutdown features. A substantially smaller version of the power transistor in parallel is a popular means of sensing current because small sensing transistor does not add any series resistance (i.e., dropout voltage) to the power path. Placing this sensor in the middle of the large power array is important to combat the mismatching forces already inherent to the pair, such as short-channel lengths (e.g., 0.18–0.35 μm) and large spreads in aspect ratios (e.g., the power device is often one-to-several thousand times larger than the sensor). This strategy often creates a discontinuity in the layout structure because one of the sensor terminals is not common to the power transistor, but the matching benefits far outweigh the resulting routing

idiosyncrasies. Although a temperature sensor is typically a different type of device, the sensor should also be close to the power device, but not necessarily in the middle, as the power device is the fundamental heat source in a linear regulator IC.

6.2 Buffer

The objective of the buffer is to drive the power pass device quickly (i.e., with high bandwidth BW_{BUF}) and with minimal power (i.e., low quiescent current I_Q). Because the large pass switch necessarily presents large parasitic capacitance C_{SW}, the buffer must not only produce low output impedance Z_{BUF} but also sink and source sufficiently high slew-rate currents I_{SR} to quickly charge and discharge C_{SW}. The buffer must also be able to swing its output voltage v_{BUF} wide enough to both shut off and fully engage the power device (i.e., $\Delta v_{BUF(max)}$ should be high), and do so under low input supplies V_{IN}, since the power lost across the power pass device decreases with decreasing V_{IN}. Additionally, the buffer should exhibit high-resistance (i.e., $R_{I.BUF}$) and low-capacitance (i.e., $C_{I.BUF}$) characteristics at its input to decrease their loading effects on error amplifier A_{EA}, which limit the regulator's loop gain and bandwidth. In summary, BW_{BUF} must be high, Z_{BUF} low, I_Q low, I_{SR} high, and $\Delta v_{BUF(max)}$ high, and the circuit must survive low V_{IN} and present high $R_{I.BUF}$ and low $C_{I.BUF}$ to A_{EA}.

6.2.1 Driving N-Type Power Switches

NPN BJT

Among all the single-transistor configurations possible, emitter and source followers subscribe best to the high-bandwidth and low-output-impedance requirements of the buffer. Of these, given the voltage drop from V_{IN} to the output of the buffer (i.e., v_{BUF}) is in the critical dropout path of a power npn BJT, p-type followers, as shown with Q_{PBUF} and M_{PBUF} in Fig. 6.10, are best suited because their respective

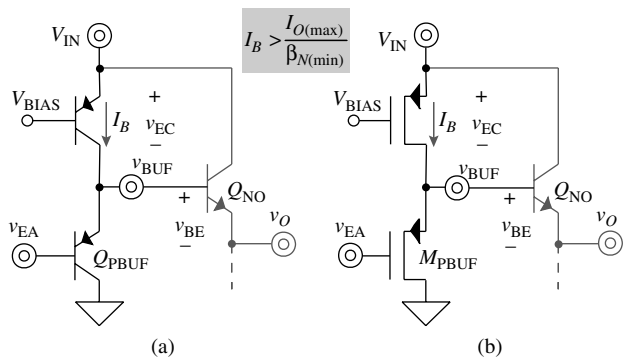


Figure 6.10 (*a*) PNP emitter-follower Q_{PBUF} and (*b*) PMOS source-follower M_{PBUF} buffers suitable to drive power npn BJTs (i.e., Q_{NO}).

emitter-base and source-gate voltages do not constrain dropout voltage V_{DO}. Despite the seemingly attractive feature of simplicity, p-type followers suffer from poor power efficiency because biasing current source I_B must be sufficiently large (even during light loading conditions) to supply the base current necessary to drive the power transistor during worst-case loading, temperature, and process-corner conditions, which translates to $I_{O(\max)}/\beta_{N(\min)}$ (at cold temperatures and weak process corners).

Designing a dynamically adaptive bias current is probably the best approach for improving efficiency during light loading conditions, which is another way of saying supplying large base currents only when demanded. Adding a negative-feedback path from the following buffer to the sourcing current, as shown in Fig. 6.11, achieves this, as the feedback loop supplies whatever sourcing current i_{BUF} is necessary to ensure v_{BUF} follows A_{EA}'s output v_{EA} by one constant V_{SG} (or V_{EB}) dc voltage. Technically, the loop is series-mixed and shunt-sampled because M_{PBUF} mixes v_{EA} and v_{BUF} with its source and gate and samples v_{BUF} with its source, which means the buffer's output impedance is that of the follower divided by the loop gain of the negative-feedback circuit (i.e., LG_{FB}):

$$R_{BUF} \approx \frac{1}{g_{m.BUF}LG_{FB}} \approx \frac{1}{g_{m.BUF}(r_{ds4}g_{m2})} \tag{6.3}$$

Note that replacing NMOS and PMOS FETs with npn and pnp BJTs, and vice versa, yields similar benefits, albeit at slightly different current and impedance levels.

Transistors M_{NB3} and M_{PB4} are current-biasing transistors with the former sinking more current the latter sources (i.e., $I_{B3} > I_{B4}$) to use the difference (i.e., $I_{B3} - I_{B4}$) to bias M_{PBUF}. Common-base current buffer

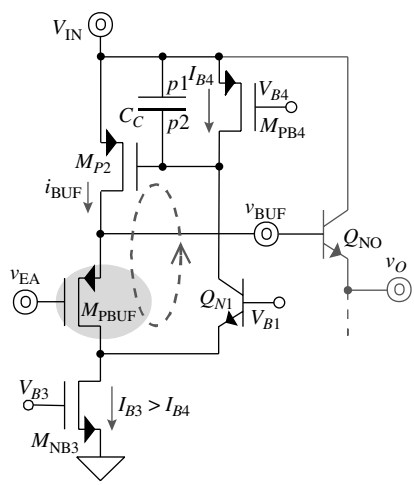


FIGURE 6.11 Dynamically adaptive PMOS source-follower buffer M_{PBUF} in negative feedback with sourcing transistor M_{P2} designed to drive power npn BJTs (i.e., Q_{NO}).

Q_{N1} folds the ac current flowing down M_{PBUF} up to M_{PB4}, where r_{ds4} converts the current into a voltage and applies the voltage to the gate of common-source amplifier M_{P2}. Although the base of the power transistor shares its connection with M_{PBUF}'s source, which has low resistance, large parasitic switch capacitance C_{SW} pulls the pole at v_{BUF} to low frequencies, near the pole the gate of sourcing transistor M_{P2} is present. The purpose of compensation capacitor C_C is to therefore pull the latter pole further down to lower frequencies and ensure it remains the dominant low-frequency pole of the negative-feedback loop.

Capacitor C_C's first-level polysilicon plate or poly 1, for short, is connected to a low-impedance source (e.g., V_{IN}) to shunt substrate noise coupled through the poly-1 terminal to ac ground, as this plate is closest to substrate and more prone to receiving noise (i.e., to substrate-noise injection). Incidentally, the capacitor is tied to V_{IN} (instead of ground) to ensure any noise in the supply couples to M_{P2}'s gate so that supply noise remains common mode (i.e., in phase) with respect to M_{P2}'s source for better power-supply rejection. M_{P2}'s gate drive, which depends on V_{IN} (as M_{P2}'s source voltage) and Q_{N1} (and its biasing voltage V_{B1}), and aspect ratio ultimately determine the maximum current the circuit can drive to supply the power transistor's worst-case base current:

$$I_{\text{BUF}} \geq 0.5K_P' \left(\frac{W}{L}\right)(V_{\text{SG2(min)}} - V_{\text{TP}})^2 \geq \frac{I_{O(\max)}}{\beta_{N(\min)}} \tag{6.4}$$

However, under light loading conditions, the quiescent current of the circuit can be as low as a few microamps. The tradeoff for this power-efficiency benefit is a slower response time because the negative-feedback loop, as evidenced by the presence of additional capacitance C_C, necessarily decreases the bandwidth of the otherwise open-loop source follower.

NMOSFET

The power NMOS device has similar requirements as its npn counterpart, except driving a base current is no longer necessary, which means the requirements of the p-type follower buffers shown in Fig. 6.10 are less stringent. Bias current I_B must now supply and sink the slew-rate current ($I_{\text{SR}\pm}$) necessary to charge and discharge large parasitic switch capacitor C_{SW}, which is not trivial but often easier than supplying a worst-case base current:

$$I_B \geq C_{\text{SW}}\left(\frac{\Delta v_C}{\Delta t_C}\right) \approx C_{\text{SW}}\Delta v_{\text{BUF(max)}}\text{BW}_{\text{BUF(min)}} \equiv I_{\text{SR}\pm} \tag{6.5}$$

where Δt_C is the time required for the voltage across C_{SW} to traverse Δv_C in response to a load dump and $\Delta v_{BUF(max)}$ and $BW_{BUF(min)}$ (i.e., $\Delta t_{C(max)}^{-1}$) represent the maximum Δv_C required and maximum Δt_C allowed. Note p-type follower Q_{PBUF} or M_{PBUF}, with enough base or gate drive, can more easily sink negative slew-rate current I_{SR-}. Again, as with the npn BJT, adding a negative-feedback loop (Fig. 6.11) can reduce the quiescent current of the circuit by dynamically adjusting I_B in Fig. 6.10 to supply maximum current only during a transient event, but like before, efficiency improves at the cost of speed.

6.2.2 Driving P-Type Power Switches

PNP BJT

The electrical characteristics of emitter and source followers synergize with the requirements of the driving buffer, but of these, p-type followers (Fig. 6.12) are often more practical for driving p-type power transistors because their outputs rise sufficiently high (e.g., within v_{EC} or v_{SD} of V_{IN}) to reliably shut off the power device. Only raising v_{BUF} within a base-emitter or gate-source voltage of V_{IN}, which is what n-type followers essentially achieve, is insufficiently high to ensure a substantially large and leaky power transistor is off across process and temperature corners during light loading conditions, where any undesired leakage current constitutes unacceptably high quiescent losses. The other major dc requirement of the buffer is to sink the worst-case base current of the pnp BJT, which amounts to $I_{O(max)}/\beta_{P(min)}$.

The challenges of sinking a large base current are the driving limits of error amplifier A_{EA}. In the case of a pnp follower buffer, for instance, as shown in Fig. 6.12*a*, sinking $I_{O(max)}/\beta_{P(min)}^2$ (e.g., 100 mA/15^2 or 444 μA) presents difficulties for A_{EA} when only allowed to operate with a minimal quiescent current, such as 10 μA or less. Similarly, a moderately sized PMOS follower (e.g., M_{PBUF} in Fig. 6.12*b*) sinking

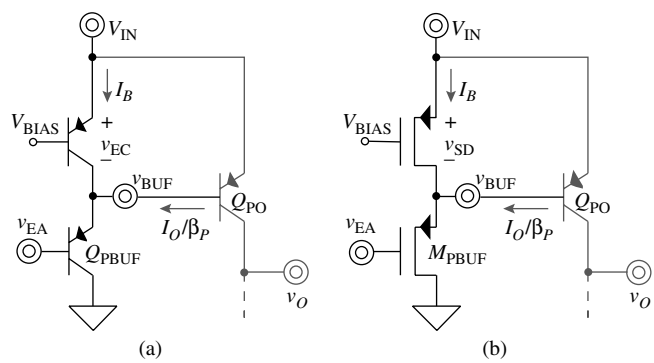


Figure 6.12 (*a*) PNP emitter- and (*b*) PMOS source-follower buffers suitable for power pnp BJT applications.

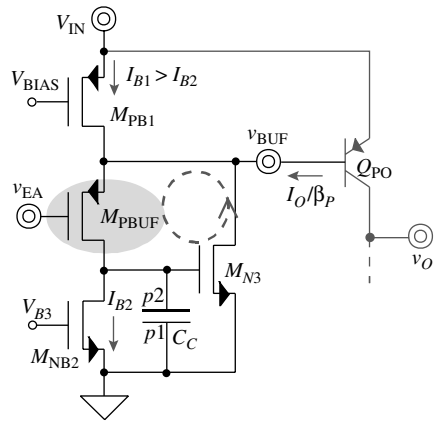


Figure 6.13 Dynamically adaptive PMOS source-follower buffer M_{PBUF} in negative feedback with current-sinking transistor M_{N3} designed to drive power pnp BJTs (i.e., Q_{PO}).

$I_{O(max)}/\beta_{P(min)}$ (e.g., 100 mA/15 or 6.7 mA) requires gate-drive voltages on the order of 1.5–3 V or higher (i.e., $V_{SG.PBUF} \approx |V_T| + [2I_D/K_P'(W/L)]^{1/2}$), which are difficult to achieve under low V_{IN} (i.e., headroom) conditions because $V_{IN(min)}$ must exceed $V_{EB.PO(max)} + V_{SG.PBUF(max)}$ for the circuit to operate. Alternatively, a negative-feedback loop placed around sinking follower buffer M_{PBUF} to another sinking device (e.g., M_{N3} in Fig. 6.13) can supplement additional base current only when required.

The idea is for feedback transistor M_{N3} (in the buffer driver of Fig. 6.13) to supplement PMOS follower M_{PBUF} with substantial sinking current when heavily loaded and use little dc current when output current I_O is near zero. The fact is M_{PBUF} only sinks nominal biasing current I_{B2} and M_{N3} sinks what remains of I_{B1} (i.e., I_{N3} is $I_{B1} - I_{B2}$) when power pnp transistor Q_{PO} is off. As the load (and the demand for Q_{PO}'s base current) increases, v_{EA} and subsequently v_{BUF} decrease, and M_{N3} sinks whatever additional current is necessary to ensure v_{BUF} is a constant V_{SG} drive above v_{EA} (as though V_{SG} were a dc offset or a battery voltage). As before, compensation capacitor C_C ensures the dominant low-frequency pole of the loop is at the gate of M_{N3} and connecting C_C to ground through its poly-1 plate ensures ground-bounce noise in M_{N3}'s source remains common mode with M_{N3}'s gate (i.e., in phase), directing in the process substrate-coupled noise to ground. In the end, as before, the quiescent current saved in employing negative feedback to adapt A_{BUF} dynamically to Q_{PO} offsets the slower response that results from compensating the negative-feedback loop.

Although not always an option, more sophisticated power biCMOS technologies offer the necessary mask set and processing steps to isolate a low threshold voltage NMOS transistor, rendering the otherwise unattractive NMOS follower practical in driving p-type power transistors, now that it can raise its output within a smaller margin of V_{IN}. Substrate NMOS transistors that do not undergo the

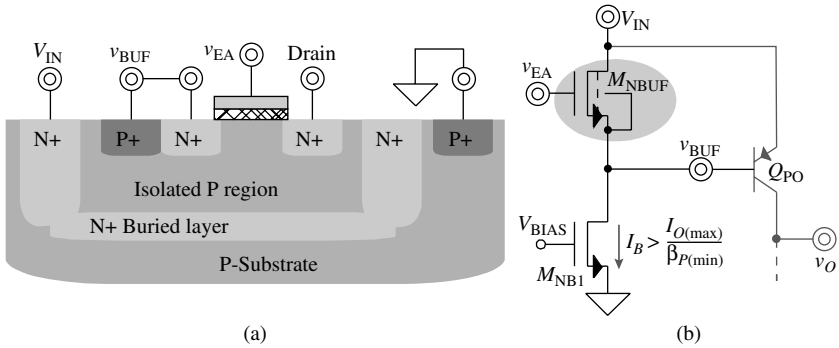


FIGURE 6.14 (*a*) Physical profile view of an isolated natural (or native) substrate NMOSFET and (*b*) its application to n-type follower buffer M_{NBUF} when driving power pnp BJT Q_{PO}.

threshold-adjustment process, for example, known as *natural* or *native* NMOSFETs, typically exhibit zero bulk-bias threshold voltages V_{TN0} near 0 V. What is more, isolating the bulk or body of this device by effectively placing the transistor in its own well and short-circuiting its bulk and source terminals avoids bulk effects and consequently incurs no impact on threshold voltage V_{TN}, reducing gate-source voltage V_{GS} to a mere $V_{DS(sat)}$ (i.e., $V_{GS} = V_{TN} + V_{DS(sat)} = V_{TN0} + V_{DS(sat)} \approx V_{DS(sat)}$). Figure 6.14*a* illustrates graphically how a deep n+ ring and an overlapping n+ buried layer isolates the aforementioned p-type bulk region from the substrate.

The isolated natural NMOS transistor, when used in a follower configuration to drive a power pnp device, however, as shown in Fig. 6.14*b*, does not, on its own, overcome the need for sinking a large base current, so a large biasing current I_B must exist. As before, though, applying negative feedback around the M_{NBUF} follower to a sinking transistor can decrease quiescent current overhead by sinking base current only when needed. In the illustrative example shown in Fig. 6.15*a*, buffer transistor M_{NBUF} and transconductor amplifier G_{FB} mix v_{EA} and v_{BUF} and V_{REF} and v_B, respectively, which means M_{N1} sources whatever current is necessary to ensure v_{BUF} follows v_{EA} with a constant but relatively small dc voltage $V_{GS.NBUF}$ and v_B equals V_{REF}.

From a biasing perspective, G_{FB} ensures the voltage across R_B is constant and equal to the difference in voltages V_{IN} and V_{REF}, which in turn sets M_{NBUF}'s biasing current. Feedback transistor M_{N1} sinks this current and whatever current power switch Q_{PO} requires when confronted with a load. Capacitor C_C sets the dominant low-frequency pole of the loop at the gate of M_{N1} and its poly-1 plate connection channels substrate noise to ground and ensures ground noise remains common mode across M_{N1}'s source and gate terminals.

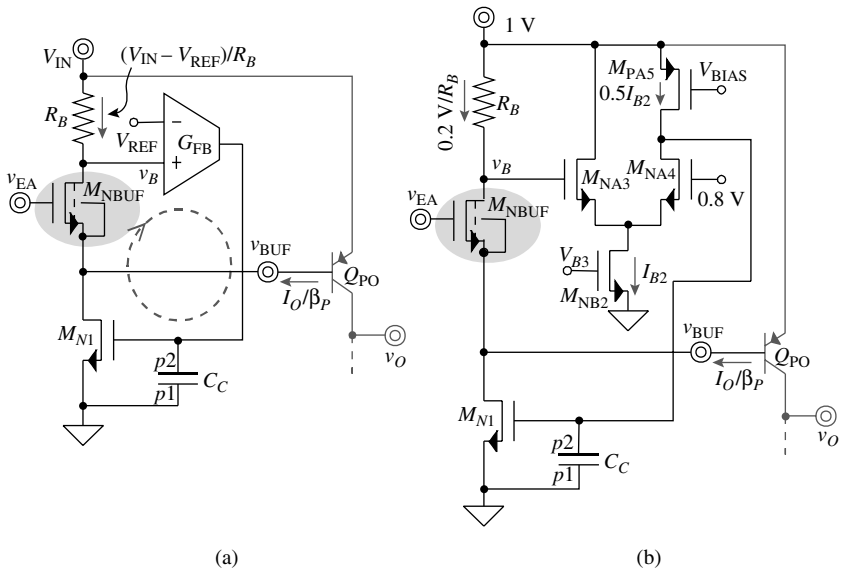


FIGURE 6.15 (*a*) Block- and (*b*) transistor-level schematic of a dynamically adaptive natural n-type follower buffer M_{NBUF} in negative feedback with current-sinking transistor M_{N1}.

Because v_{BUF} must reach high enough to shut off the pnp transistor, the voltage across R_B should be kept low at around 100–300 mV, as shown in the sample embodiment of Fig. 6.15*b* where differential pair M_{NA3}–M_{NA4} and load M_{PA5} constitute G_{FB} and V_{IN}, and V_{REF} are 1 V and 0.8 V, respectively.

PMOSFET

Like the pnp device, a p-type MOS transistor requires a high maximum input voltage $V_{BUF(max)}$ during zero-to-no load conditions to force the power device into its off state, which is why the p-type followers shown in Fig. 6.12 also work in power PMOS applications. The only problem with a p-type follower approach, however, is poor dropout performance under low V_{IN} (i.e., high-efficiency) conditions. To be more specific, dropout voltage V_{DO} in the case of a power PMOSFET M_{PO} is inversely proportional to M_{PO}'s gate drive $V_{SG.O}$ and therefore limited to the difference in supply V_{IN} and at least one source-gate or emitter-base voltage V_{SG} or V_{EB}, that is to say,

$$V_{DO} = I_O R_{DO} \propto \frac{1}{V_{SG.O}}\bigg|_{PBUF} \geq \frac{1}{V_{IN} - V_{ON.PBUF} - V_{EA(min)}} \tag{6.6}$$

where R_{DO} is M_{PO}'s dropout resistance, $V_{ON.PBUF}$ represents the source-gate or emitter-base voltage of the buffer, and $V_{EA(min)}$ is A_{EA}'s minimum

output voltage, which is often one $V_{DS(sat)}$ or $V_{CE(min)}$ above ground. Natural n-type follower buffer M_{NBUF} in Fig. 6.14, assuming an isolated natural NMOSFET is available, is more attractive than a p-type follower because it cannot only shut off the power PMOS transistor but also extend its gate drive during dropout conditions by roughly one source-gate or emitter-base voltage:

$$V_{DO} = I_O R_{DO} \propto \frac{1}{V_{SG.O}}\bigg|_{NBUF} > \frac{1}{V_{IN} - V_{EA(min)}} \tag{6.7}$$

As in the NMOS power transistor case, driving a base current may no longer be an issue but charging and discharging the large parasitic capacitor present at the gate of the power PFET is. In this respect, follower buffers can supply substantial slew-rate currents in one direction but not in the other because biasing current I_B bounds them. Applying negative feedback, as before, can supplement the driving deficiencies of the follower at a fraction of the cost in quiescent current, as exemplified by the n-type follower M_{NBUF} embodiment of Fig. 6.15 and its p-type M_{PBUF} equivalent in Fig. 6.16. Note resistor R_B must be low enough to ensure the voltage across its terminals

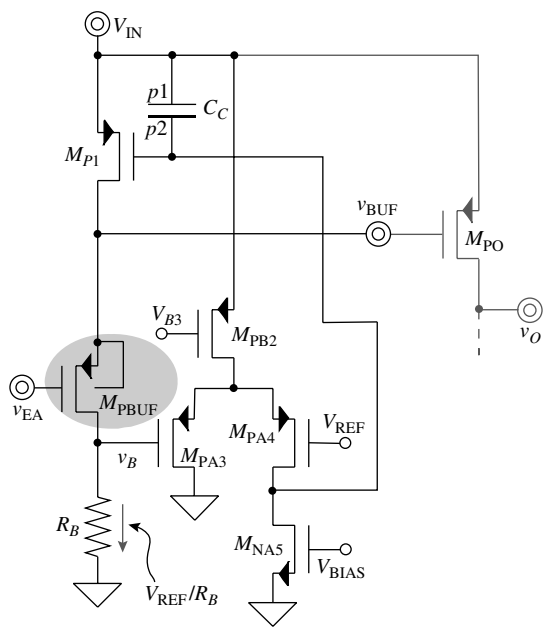


Figure 6.16 Dynamically adaptive p-type follower buffer M_{PBUF} in negative feedback with current-sourcing transistor M_{P1}.

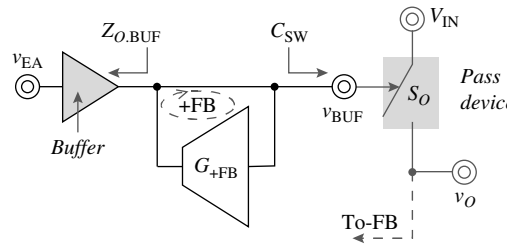


Figure 6.17 Dynamically adaptive regenerative (i.e., positive feedback) transconductance G_{+FB} buffer.

does not push buffer transistor M_{NBUF} or M_{PBUF} into triode during slew-rate conditions, which would otherwise increase $V_{GS.NBUF}$ or $V_{SG.PBUF}$ and decrease $V_{BUF(max)}$. Also, as with the others, applying negative feedback necessarily slows the circuit because the loop introduces another pole and requires its location relative to the pole at v_{BUF} to be low enough to ensure stable operating conditions.

The benefits of dynamically adapting the driving capabilities of the buffer with follower transistors in negative feedback may not always sufficiently offset their bandwidth limitations to make them attractive solutions, especially in system-on-chip (SoC) applications where output capacitor C_O is not large enough to keep the regulator's output (i.e., v_O) from drooping considerably during large and fast load-dump events. A localized regenerative circuit (in place of a dynamically biased transistor in negative feedback), as graphically depicted in Fig. 6.17, can dynamically adapt and bootstrap the otherwise slew-rate limited buffer without surrendering to the bandwidth limitation a negative-feedback loop imposes. The main issues with positive feedback are proneness to sustained oscillations and difficult-to-reverse latched events, which is why positive-feedback loop gain LG_{+FB} must remain below unity:

$$LG_{+FB} \approx G_{+FB}Z_{O.BUF} \approx \frac{G_{+FB}}{g_{m.BUF}LG_{REG}\left(\frac{g_{m.BUF}LG_{REG}s}{C_{SW}}+1\right)} < 1 \qquad (6.8)$$

where G_{+FB} is the effective transconductance gain of the positive-feedback path, $Z_{O.BUF}$ the output impedance of the buffer, LG_{REG} the loop gain of the regulator (i.e., of the outer negative-feedback loop, the one regulating v_O), and C_{SW} the parasitic filter capacitance at v_{BUF}. Note one of the effects of the loop that regulates v_O about V_{REF} (i.e., LG_{REG}) is to decrease the buffer's open-loop output impedance because S_O's input terminal essentially shunt-samples v_{BUF}. As a result, ensuring LG_{+FB} is less than 1 V/V is equivalent to saying LG_{REG} exceeds LG_{+FB} at high frequencies, when LG_{REG} is less than 1 V/V and

its shunting effect on $1/g_{m.BUF}$ is therefore negligible, which means $Z_{O.BUF}$ is roughly $1/g_{m.BUF}$:

$$LG_{REG} \approx A_{EA}A_{PO}\beta_{FB} = \frac{A_{EA}A_{PO}R_{FB1}}{R_{FB1}+R_{FB2}} > LG_{+FB}\big|_{LG_{REG}<1} \approx \frac{G_{+FB}}{g_{m.BUF}\left(\frac{g_{m.BUF}s}{C_{SW}}+1\right)} \tag{6.9}$$

where A_{PO} refers to the voltage gain across the power device and β_{FB} the feedback factor of the outer loop (Fig. 6.1). In more qualitative terms, the output impedance of the follower buffer $Z_{O.BUF}$ (in negative feedback with power switch S_O through the loop that regulates v_O) must be low enough to damp the oscillations and latching tendencies that result from the regenerative (i.e., positive-feedback) loop.

When considering the biasing condition of the follower buffer, it is worthwhile to recall that the buffer's bandwidth requirement varies with output current I_O, which is why externally compensated regulators (where output pole p_O is dominant) conventionally place buffer pole p_{BUF} (which resides at node v_{BUF}) at its worst-case high-frequency location, when I_O and unity-gain frequency f_{0dB} are highest. The problem is output pole p_O, as graphically illustrated in Fig. 6.18, increases with I_O (i.e., p_O is proportional to r_o^{-1} and therefore proportional to I_O), which means f_{0dB} also increases with I_O, and worst-case extremes of R_{ESR} values exacerbate the total resulting variation. Ensuring p_{BUF} does not fall far below f_{0dB} (i.e., p_{BUF} remains above $p_{BUF(min)}$) translates to biasing the follower buffer with enough quiescent current to guarantee a sufficiently low buffer impedance $Z_{O.BUF}$ (in the presence of a large C_{SW}) across process and temperature corners.

In the same spirit of dynamic adjustments, considering high efficiency is paramount in battery-powered applications, having the regenerative loop presented in Fig. 6.17 also include a dc component

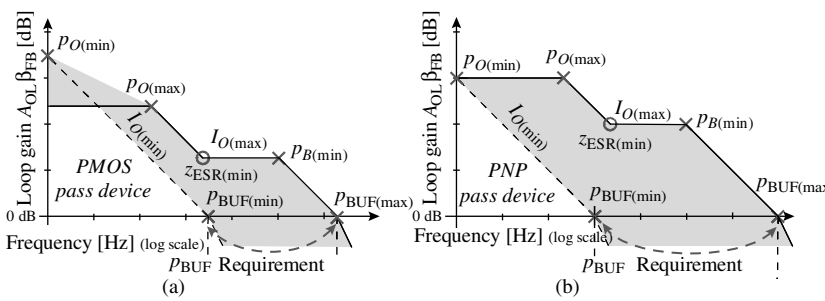


FIGURE 6.18 Worst-case buffer pole p_{BUF} requirements with respect to output current I_o and the zero-pole pair that equivalent series resistor R_{ESR} introduces (i.e., z_{ESR} and p_B) in externally compensated regulators with power p-type (*a*) MOSFETs and (*b*) BJTs.

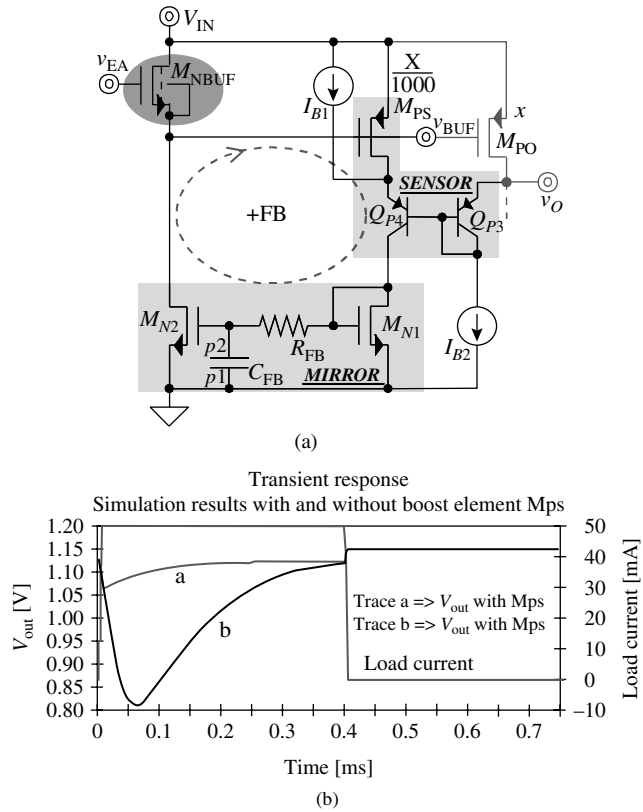


FIGURE 6.19 (*a*) Dynamically adaptive pole-tracking natural n-type follower buffer M_{NBUF} in positive feedback with current-sinking transistor M_{N1} and (*b*) its transient-response performance.

whose net result was to increase the biasing current of the buffer with I_O would be ideal. The circuit illustrated in Fig. 6.19*a* does just this, employ positive feedback to accelerate the transient response of v_{BUF} while simultaneously adjusting the steady-state biasing conditions of follower buffer M_{NBUF}. In short, sensing circuit M_{PS}–Q_{P3}–Q_{P4} sources a bias current into folding mirror M_{N1}–M_{N2} that is proportional to I_O, establishing in the process a bias current for M_{NBUF} that increases with I_O. Figure 6.19*b* shows how the positive-feedback path accelerates the time the regulator requires to respond to a 0–5-mA, 100-ns load dump from roughly 60 μs (without positive-feedback transistor M_{PS}) to less than 3 μs (with M_{PS}) with similar zero-biasing conditions.

To be more explicit, substantially smaller sensing PMOS transistor M_{PS} directs a considerably smaller fraction of I_O through current

buffer Q_{P4} to current mirror M_{N1}–M_{N2}, whose output current biases M_{NBUF} with respect to I_O. Source current I_{B1} ultimately biases M_{NBUF} during zero-to-light loading conditions (i.e., when I_O is low) because mirror M_{N1}–M_{N2} sinks a proportional ratio of I_{B1} and M_{PS}'s drain current I_{PS}, and I_{PS} is negligibly small when the regulator is lightly loaded. Note that pulling a bias current directly from v_{BUF}, altogether bypassing current buffer Q_{P4} and mirror M_{N1}–M_{N2}, to bias M_{NBUF} during light loads would starve buffer-mirror transistors Q_{P4}, M_{N1}, and M_{N2} of current and correspondingly increase their response time (i.e., decrease their combined bandwidth). The underlying purpose of sensor buffer Q_{P3}–Q_{P4} is to equate, to first order, M_{PS}'s drain voltage to that of M_{PO}'s (i.e., ensure $V_{SD.PS}$ is roughly $V_{SD.PO}$) so as to prevent M_{PS} from entering a dissimilar region of operation with respect to M_{PO}. This "fix" is important during dropout conditions, when M_{PO} enters the triode region and v_{BUF} drops close to ground, without which M_{PS} would remain in saturation and be induced to source a disproportionately larger current into mirror M_{N1}–M_{N2}.

During a fast positive load dump, when I_O suddenly increases and A_{EA} responds by decreasing v_{EA}, v_{BUF} drops and, as it does, M_{PS} drives more current into the current mirror so M_{N2} pulls more current from v_{BUF} and increases the discharge rate of parasitic capacitor C_{SW} and the falling slew rate of v_{BUF}. Increasing the aspect ratio of M_{PS} and the mirror gain of M_{N1}–M_{N2} increases the transconductance gain of this positive-feedback loop (i.e., increases G_{+FB}) and magnifies its "speed-up" effects, except larger values may compromise the stability performance of the regulator. As a result, the aspect ratios of sense PFET M_{PS} and mirror transistor M_{N2} with respect to M_{N1} are optimum when they are relatively large, but just below the point they induce unacceptably long and/or large oscillations in any one of the worst-case process, temperature, and operating corners of the system. This worst-case point may fall in one of several corners, such as with strong PMOSFETs (e.g., strong M_{PS} and M_{PO}), weak NMOSFETs (e.g., weak M_{NBUF}), low temperature (i.e., stronger FETs), high V_{IN} (e.g., more gate drive and therefore stronger M_{PS} and M_{PO}), high I_O and low C_O (e.g., high p_O), and high R_{ESR} (e.g., low z_{ESR}).

Placing low-pass filter R_{FB}-C_{FB} in the positive-feedback path as shown helps decouple the dc and ac performance of the positive-feedback loop because decreasing LG_{+FB} sooner in frequency (for stability) affords the circuit more margin for higher low-frequency positive-feedback gain. In practical terms, the filter extends the biasing range of M_{NBUF} (i.e., $I_{NBUF(min)}$ to $I_{NBUF(max)}$) across I_O by offsetting the larger M_{N1}–M_{N2} mirror gain and the larger positive-feedback gain it implies with a slower positive-feedback response. In other words, the filter decreases LG_{+FB} at higher frequencies, ensuring LG_{+FB} falls below LG_{REG} across all frequencies of interest, where LG_{REG} decreases and nears f_{0dB}.

Circuit Parameter	Specification	PMOS Process Parameter	Value
V_{IN}	1.1–1.6 V	$\lvert V_{TP}\rvert$ and V_{TN}	0.6 V ± 150 mV
I_0	≤ 50 mA	$V_{TN.NAT}$	0 V ± 150 mV
BW_0	≥ 100 kHz	K'_P	40 μA/V² ± 20%
BW_{50mA}	≥ 1 MHz	K'_N	100 μA/V² ± 20%
$I_{Q.0}$	≤ 4 μA	L	≥ 0.35 μm
$V_{BUF(max)}$	$V_{IN} - \lvert V_{TP(min)}\rvert$	$\lambda_{L(min)}$	(500 mV)$^{-1}$
$V_{BUF(min)}$	0.2 V	L_{OV}	35 nm
		C''_{OX}	2.5 fF/μm²
		$\beta_{0.PNP}$	50–150 A/A
		$V_{A.PNP}$	15 V

TABLE 6.3 Target Specifications and Process-Parameter Values for the Buffer in Example 6.2

Design Example 6.2 Refer to the specifications and process parameters outlined in Table 6.3 and design the dynamically adaptive, pole-tracking positive-feedback buffer illustrated in Fig. 6.19 to drive the power pass device designed in Example 6.1. The no- and full-load (i.e., 50 mA) bandwidths of the circuit should be higher than 100 kHz and 1 MHz, respectively, with a no-load quiescent current of less than 4 μA and an output voltage swing ranging from 0.2 V to $V_{IN} - \lvert V_{TP(min)}\rvert$ (to shut off power PMOS M_{PO}). A 0.9–1.6-V NiMH battery supplies the buffer but its intended range is only 1.1–1.6 V because (1) MOSFETs with 0.6 V ± 150 mV thresholds in analog circuits constrain headroom limit $V_{IN(min)}$ to $\lvert V_{T(max)}\rvert + 2V_{DS(sat)}$ (e.g., 1–1.1 V) and (2) the energy that remains in the battery across its 0.9–1 V range is considerably low.

Architecture notes

1. Common-base transistor Q_{P4}:

   At $V_O = 1$ V and $V_{GS.N1} \approx 1$ V, Q_{P4} is saturated:

   $$V_{EC.P4} \approx (V_O - V_{EB.P3} + V_{EB.P4}) - V_{GS.N1} \approx V_O - V_{GS.N1} \approx 1\text{ V} - 1\text{ V} = 0$$

   so shift I_{B1} and insert dc level-shifting resistor R_{DC} as shown in Fig. 6.20:

   $$V_{EC.P4} \approx V_O - (V_{GS.N1} - I_{B1}R_{DC}) \geq 300\text{ mV}$$

   ∴ Q_{P4} is now only lightly saturated (i.e., still in a high-gain mode).

Transistor design

2. Zero-load parasitic capacitance $C_{BUF.0}$ (at v_{BUF}):

   $$C_{BUF.0} \approx C_{SG.PO} \approx \frac{2}{3}C''_{OX}W_{PO}L_{PO} + C''_{OX}W_{PO}L_{OV}$$
   $$\approx \frac{2}{3}(5\text{ f})(18.3\text{ k})(0.35) + (5\text{ f})(18.3\text{ k})(0.035) \approx 25\text{ pF}$$

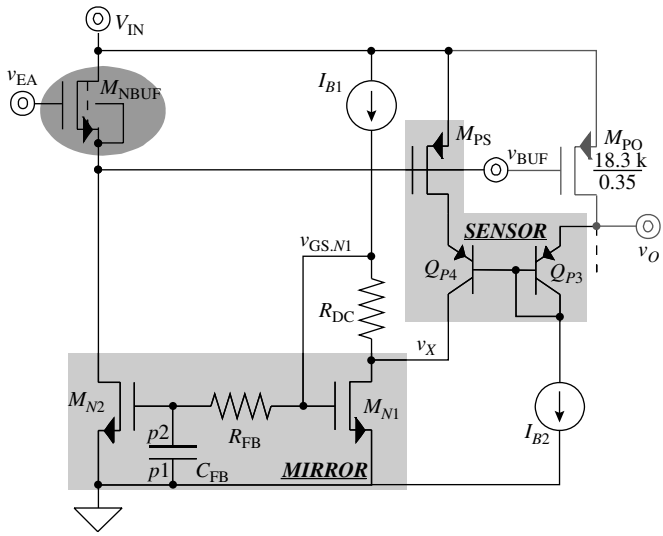


FIGURE 6.20 Dynamically adaptive natural n-type follower buffer M_{NBUF} in positive feedback with current-sinking transistor M_{N1} and modified to accommodate an output voltage of 1 V (and prevent common-base transistor Q_{P4} from deeply saturating).

3. Full-load parasitic capacitance $C_{BUF.50mA}$ (at v_{BUF}):

$$C_{BUF.50mA} \approx C_{SG.PO} + C_{DG.PO(Miller)}$$

$$\approx \frac{2}{3}C''_{OX}W_{PO}L_{PO} + C''_{OX}W_{PO}L_{OV}(g_{m.PO}r_{sd.PO})$$

$$\approx \frac{2}{3}C_{OX}W_{PO}L_{PO} + C_{OX}W_{PO}L_{OV}\left(\frac{\sqrt{2I_{O(max)}K'_P\left(\frac{W}{L}\right)_{PO}}}{I_{O(max)}\lambda_{L(min)}}\right)$$

$$\approx \frac{2}{3}(5\text{f})(18.3\text{k})(0.35) + (5\text{f})(18.3\text{k})(0.035)\left(\frac{\sqrt{2(50\text{m})(40\mu)\left(\frac{18.3\text{k}}{0.35}\right)}}{(50\text{m})(500\text{m})}\right)$$

$$\approx 80\text{ pF}$$

4. Full-load bandwidth BW_{50mA}:

$$BW_{50\,mA} \approx \frac{1}{2\pi C_{BUF.50\,mA}R_{BUF}}$$

$$\approx \frac{g_{m.NBUF}}{2\pi C_{BUF.50\,mA}} = \frac{\sqrt{2I_{NBUF.50\,mA}K'_N\left(\frac{W}{L}\right)_{NBUF}}}{2\pi C_{BUF.50\,mA}} \geq 1\,\text{MHz}$$

$$\text{or } I_{\text{NBUF.50mA}}\left(\frac{W}{L}\right)_{\text{NBUF}} \geq \frac{\left[2\pi C_{\text{BUF.50mA}}\text{BW}_{\text{50mA}}\right]^2}{2K'_{N(\text{min})}} = \frac{\left[(2\pi)(80\,\text{p})(1\text{MHz})\right]^2}{2(80\,\mu)} \approx 1.6\,\text{m}$$

$$\therefore \text{ if } \left(\frac{W}{L}\right)_{\text{NBUF}} \equiv \frac{12\,\mu\text{m}}{0.35\,\mu\text{m}} \quad \text{then} \quad I_{\text{NBUF.50mA}} \approx 46\,\mu\text{A}$$

(Note a smaller M_{NBUF} implies a higher $\text{LG}_{+\text{FB}}$ and a larger M_{NBUF} presents a larger parasitic capacitance to A_{EA}.) Choosing a 1:1 M_{N1}–M_{N2} mirror ratio means $(W/L)_{N1}$ equals $(W/L)_{N2}$ by design

$$\text{and} \quad \frac{\left(\frac{W}{L}\right)_{\text{PO}}}{\left(\frac{W}{L}\right)_{\text{PS}}} = \frac{\left(\frac{18.3\,\text{k}}{0.35}\right)}{\left(\frac{W}{L}\right)_{\text{PS}}} \equiv \frac{50\,\text{mA}}{46\,\mu\text{A}}$$

$$\therefore \left(\frac{W}{L}\right)_{\text{PS}} \geq 48 \quad \text{or} \quad \left(\frac{W}{L}\right)_{\text{PS}} \equiv \frac{17\,\mu\text{m}}{0.35\,\mu\text{m}}$$

5. Zero-load bandwidth BW_0:

$$\text{BW}_0 \approx \frac{1}{2\pi C_{\text{BUF}} R_{\text{BUF}}} \approx \frac{g_{m.\text{NBUF}}}{2\pi C_{\text{BUF.0}}} = \frac{\sqrt{2I_{\text{NBUF.0}}K'_N\left(\frac{W}{L}\right)_{\text{NBUF}}}}{2\pi C_{\text{BUF.0}}} \geq 100\,\text{kHz}$$

$$\text{or} \quad I_{\text{NBUF.0}} \geq \frac{\left[2\pi C_{\text{BUF.0}}\text{BW}_0\right]^2}{2K'_{N(\text{min})}\left(\frac{W}{L}\right)_{\text{NBUF}}} = \frac{\left[2\pi(25\,\text{p})(100\,\text{kHz})\right]^2}{2(80\,\mu)\left(\frac{12}{0.35}\right)} \approx 0.05\,\mu$$

$\therefore$ (while adding margin to swamp coupled noise) $I_{\text{NBUF.0}} \equiv 0.5\,\mu\text{A}$.

6. $V_{\text{BUF(min)}}$:

$$V_{\text{BUF(min)}} \approx V_{\text{DS.}N2\text{(sat)}} \leq \sqrt{\frac{2I_{\text{NBUF(max)}}}{K'_{N(\text{min})}\left(\frac{W}{L}\right)_{N2}}} \leq 0.2\,\text{V}$$

so choosing $3\text{L}_{(\text{min})}$ to minimize channel-length modulation (i.e., λ) effects:

$$\therefore \left(\frac{W}{L}\right)_{N2} \geq \frac{2I_{\text{NBUF(max)}}}{V^2_{\text{DS.}N2\text{(sat)}}K'_{N(\text{min})}} = \frac{2(46\,\mu)}{(0.2)^2(80\,\mu)} = 28.7$$

$$\text{or} \quad \left(\frac{W}{L}\right)_{N2} \equiv \left(\frac{W}{L}\right)_{N1} \equiv \frac{30\,\mu\text{m}}{3(0.35)\,\mu\text{m}}$$

7. Positive-feedback filter R_{FB}-C_{FB}:
Set feedback-filter pole $f_{P.FB}$ to 100 kHz to ensure positive loop gain LG_{+FB} starts dropping past 100 kHz:

$$f_{P.FB} \approx \frac{1}{2\pi C_{FB} R_{FB}} \leq 100\ \text{kHz}$$

so if $C_{FB} \equiv 5$ pF (considering that a larger C_{FB}, when using 5-fF/μm² capacitors, requires more than 32 μm × 32 μm of silicon area),

$$\therefore \quad R_{FB} \geq \frac{1}{2\pi C_{FB} f_{P.FB}} = \frac{1}{2\pi(5\ \text{p})(100\ \text{k})} \approx 320\ \text{k}\Omega$$

8. Ensure Q_{P4} only saturates lightly (i.e., $V_{EC.P4} < V_{EC(min)} \approx 0.3$ V):

$$V_{EC.P4(min)} \approx V_O - (V_{GS.N1(max)} - I_{B1} R_{DC})$$

$$\approx V_O - V_{TN(max)} - \sqrt{\frac{2I_{NBUF(max)}}{K'_{N(min)}\left(\frac{W}{L}\right)_{N1}}} + I_{B1} R_{DC} \geq 0.3\ \text{V}$$

and having picked I_{B1} as 0.5 μA,

$$R_{DC} \geq \frac{V_{EC.P4(min)} - V_O + V_{TN(max)} + \sqrt{\frac{2I_{NBUF(max)}}{K'_{N(min)}\left(\frac{W}{L}\right)_{N1}}}}{I_{B1}}$$

$$\approx \frac{(0.3) - (1) + (0.75) + \sqrt{\frac{2(46\ \mu)}{(80\ \mu)\left(\frac{30}{1.05}\right)}}}{(0.5\ \mu)} = 501\ \text{k}\Omega \qquad \therefore \qquad R_{DC} \equiv 500\ \text{k}\Omega$$

9. Bias current I_{B2}:
Choosing I_{P3} as 0.5 μA by design

$$I_{B2(min)} \geq \frac{I_{P4(max)}}{\beta_{(min)}} + I_{P3} \approx \frac{(46\ \mu)}{(50)} + (0.5\ \mu) = 1.4\ \mu\text{A}$$

$\therefore$ Choosing I_{B2} as 2 μA allows a ±30% variation in I_{B2} keep the current above the 1.4 μA target.

Design checks

10. Positive-feedback stability:

a. Since loop gain LG_{+FB} (from v_{BUF} back to v_{BUF}) is 0 V/V at $I_O = 0$ mA (because M_{PS} is off)

∴ No ac positive-feedback effects exist when the regulator is unloaded and the circuit therefore remains stable, assuming the outer loop is stable to begin with.

b. At $I_O = 50$ mA, LG_{+FB} is greatest with a worst-case high-frequency (HF) gain of (when outer loop gain LG_{REG} is below 1 V/V and its effects on $1/g_{m.NBUF}$ are therefore negligible)

$$LG_{+FB} \equiv \frac{v_{buf}}{v_{buf}} = \left(\frac{v_{g.N1}}{v_{buf}}\right)\left(\frac{v_{buf}}{v_{g.N1}}\right) \approx \left(\frac{g_{m.PS}}{g_{m.N1}}\right)\left(\frac{g_{m.N2}}{g_{m.NBUF}\left(\frac{g_{m.NBUF}s}{C_{SW}}+1\right)}\right)$$

$$\approx \frac{g_{m.PS}}{g_{m.NBUF}\left(\frac{g_{m.NBUF}s}{C_{SW}}+1\right)} \leq \frac{\sqrt{2I_{PS}K_P'\left(\frac{W}{L}\right)_{PS}}}{\sqrt{2I_{PS}K_N'\left(\frac{W}{L}\right)_{NBUF}}} = \sqrt{\frac{K_P'\left(\frac{W}{L}\right)_{PS}}{K_N'\left(\frac{W}{L}\right)_{NBUF}}}$$

$$= \sqrt{\frac{(40\,\mu)\left(\frac{17}{0.35}\right)}{(100\,\mu)\left(\frac{12}{0.35}\right)}} \approx 0.75 < 1$$

∴ No sustained oscillations appear below or above $f_{P.FB}$. (Applying worst-case 3-σ extremes for K_P' and K_N' simultaneously is probabilistically unreasonable. Note $f_{P.FB}$ attenuates LG_{+FB} past 100 kHz, before and after $BW_{50\,mA}$.)

11. Closed-loop small-signal buffer gain A_{BUF} (i.e., v_{buf}/v_{ea}):

a. Decomposing M_{NBUF}'s g_m into its gate-source components, as shown in Fig. 6.21*a*, reveals the positive-feedback loop mixes M_{NBUF}'s v_{ea}-derived current i_i with M_{N2}'s feedback current i_{fb}, M_{PS}'s gate senses v_{buf}, and the resistance into M_{NBUF}'s v_{buf}-derived transconductance is $r_{ds.NBUF} \,||\, 1/g_{m.NBUF}$ or roughly $1/g_{m.NBUF}$.
∴ Open-loop forward transresistance gain $A_{R.OL}$ (i.e., v_{buf}/i_e) is $1/g_{m.NBUF}$ and A_{BUF} is v_{ea}'s translation into i_i times closed-loop transresistance gain $A_{R.CL}$ or

$$A_{BUF.DC} \equiv \left(\frac{v_{buf}}{v_{ea}}\right) = \left(\frac{i_i}{v_{ea}}\right)\left(\frac{v_{buf}}{i_i}\right) = \left(\frac{i_i}{v_{ea}}\right)A_{R.CL} = g_{m.NBUF}\left(\frac{A_{R.OL}}{1-LG_{+FB}}\right)\Bigg|_{DC}$$

$$\approx g_{m.NBUF}\left[\frac{\left(\frac{1}{g_{m.NBUF}}\right)}{1-(0.75)}\right] \approx 4$$

or dc gain $A_{BUF.DC}$ is 12 dB (as corroborated with the Spice simulation results shown in Fig. 6.21*b*).

b. $LG_{+FB.DC}$ is 0 V/V at $I_O = 0$ mA
∴ $A_{BUF.DC} = 1$ or 0 dB (because no ac positive-feedback effects exist).

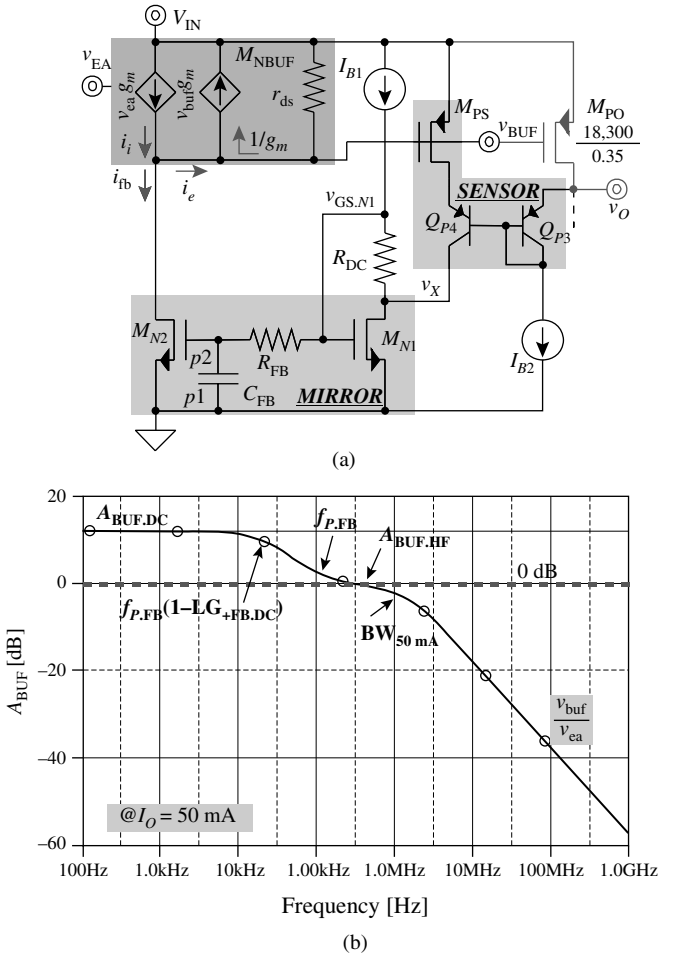


FIGURE 6.21 (*a*) Decomposing NMOS follower M_{NBUF}'s transconductance g_m into its v_{ea}– and v_{buf}-derived gate-source components to examine the positive-feedback loop and (*b*) Spice simulation results of the closed-loop small-signal buffer gain A_{BUF} or v_{buf}/v_{ea} when output current I_O is 50 mA.

c. Positive feedback converts $f_{P.FB}$ into a pole at $(1\text{-}LG_{+FB.DC})f_{P.FB}$ (i.e., $0.25f_{P.FB}$) and a zero at $f_{P.FB}$:

$$R_{CL}=\frac{R_{OL}}{1-\left[\frac{LG_{+FB.DC}}{\left(\frac{s}{2\pi f_{P.FB}}+1\right)}\right]}=\frac{\left(\frac{R_{OL}}{1-LG_{+FB.DC}}\right)\left(\frac{s}{2\pi f_{P.FB}}+1\right)}{\left[\frac{s}{2\pi\left(1-LG_{+FB.DC}\right)f_{P.FB}}+1\right]}=\frac{R_{OL.DC}\left(\frac{s}{2\pi f_{P.FB}}+1\right)}{\left[\frac{s}{2\pi\left(1-LG_{+FB.DC}\right)f_{P.FB}}+1\right]}$$

d. Since $LG_{+FB.DC}$ is less than 1 V/V, high-frequency loop gain $LG_{+FB.HF}$ and $A_{BUF.HF}$ past $f_{P.FB}$ (and before BW_{50mA}, that is, below 1 MHz) are

$$LG_{+FB.HF} = \frac{LG_{+FB.DC}}{\left(\frac{s}{f_{P.FB}}+1\right)} < \left.\frac{1}{\left(\frac{s}{f_{P.FB}}+1\right)}\right|_{f>f_{P.FB}} \ll 1$$

$$\therefore \quad A_{BUF.HF} \equiv \left.\frac{v_{buf}}{v_{ea}}\right|_{f_{P.FB}<f<BW_{50mA}} \equiv \left.\left(\frac{i_i}{v_{ea}}\right)\left(\frac{v_{buf}}{i_i}\right)\right|_{HF} \approx g_{m.NBUF}\left(\frac{R_{OL}}{1-LG_{+FB.HF}}\right)$$

$$\approx g_{m.NBUF}R_{OL} \approx \frac{g_{m.NBUF}}{g_{m.NBUF}} = 1 \qquad \text{or} \qquad 0 \text{ dB (as shown around 200 kHz in Fig. 6.21}b\text{)}$$

12. $V_{BUF(max)}$ (assuming $V_{EA(max)}$ is within one $V_{SD(sat)}$ or $V_{EC(min)}$ of V_{IN}, as for example V_{IN} – 0.3 V):

$$V_{BUF(max)} > V_{EA(max)} - V_{GS.NBUF.0(max)} = V_{EA(max)} - \left(V_{TN.NAT(max)} + \sqrt{\frac{2I_{NBUF.0}}{K'_{N(min)}\left(\frac{W}{L}\right)_{NBUF}}}\right)$$

$$\approx (V_{IN}-0.3) - \left((0.15) + \sqrt{\frac{2(0.5\,\mu)}{(80\,\mu)\left(\frac{12}{0.35}\right)}}\right) \approx V_{IN} - 0.47 \text{ V} < V_{IN} - |V_{TP(min)}|$$

∴ Absolute worst-case $V_{BUF(max)}$ is below its target specification by 20 mV (does not meet the specification).

a. Assuming ideal conditions (i.e., $V_{DS.NBUF.0(sat)}$ is nearly 0 V when lightly loaded and $V_{GS.NBUF.0(max)}$ is therefore roughly $V_{TN.NAT(max)}$):

$$V_{BUF(max)} = V_{IN} - 0.45 \text{ V}$$

∴ The circuit barely meets the specification under ideal circuit conditions; in other words, the architecture is the limiting agent.

b. Inequality $V_{BUF(max)} \geq V_{IN} - |V_{TP(min)}|$ combines 3-σ $V_{TN.NAT(max)}$ (in $V_{BUF(max)}$) and $V_{TP(min)}$ linearly, which is not realistic, so exceeding the specification by only 20 mV when applying improbable worst-case conditions implies the circuit meets the specification with an acceptable degree of risk.

13. Zero-load quiescent current $I_{Q.0}$:

$$I_{Q.0} = I_{N2.0} + I_{N1.0} + I_{P3} \approx I_{B1} + I_{B1} + I_{B2} = (0.5\,\mu\text{A}) + (0.5\,\mu\text{A}) + (2\,\mu\text{A}) = 3\,\mu\text{A}$$

so a 30% variation in $I_{Q.0}$ (i.e., $I_{Q.0}$ can be as high as 3.9 μA) keeps the buffer power within specification.

a. Full-load quiescent current $I_{Q.50mA}$ is

$$I_{Q.50mA} = I_{N2.50mA} + I_{N1.50mA} + I_{P3} \approx (46\,\mu\text{A}) + (46\,\mu\text{A}) + (2\,\mu\text{A}) = 94\,\mu\text{A}$$

which with a 30% variation, can be as high as 120 µA. Note 120 µA is still substantially below $I_{O(max)}$ (i.e., 50 mA) so its impact on full-load efficiency is negligible.

14. $V_{IN(min)}$ (assuming the minimum voltage across I_{B1} is a small $V_{SD(sat)}$, like for example 150 mV):

$$V_{IN(min)} = V_{GS.N1(max)} + V_{B1(min)} = \left(V_{TN(max)} + \sqrt{\frac{2I_{NBUF(max)}}{K'_{N(min)}\left(\frac{W}{L}\right)_{N1}}} \right) + V_{B1(min)}$$

$$\approx (0.75) + \sqrt{\frac{2(46\,\mu)}{(80\,\mu)\left(\frac{30}{1.05}\right)}} + (0.15) \approx 1.1\ \text{V}$$

∴ The circuit meets the headroom limit specified.

6.2.3 Layout

Perhaps the most important layout consideration for buffer A_{BUF} is its placement relative to error amplifier A_{EA} and power switch S_O because parasitic delays caused by physical separation decrease bandwidth performance. As such, the silicon distance between A_{EA}'s output and A_{BUF}'s input should be as short as possible, as should the separation between A_{BUF}'s output and S_O's input. Similarly, to increase the bandwidth at every point in the circuit, the layout should be generally small and modular so that its constituent paths remain short and low impedance (i.e., they introduce low parasitic series resistance and shunt capacitance).

Although A_{BUF} does not drive the currents power device S_O sources, A_{BUF} supplies considerably higher currents than A_{EA}, especially during heavy loading conditions, which is why supply rails V_{IN} and ground should be relatively wide (e.g., 5–8 µm) and fairly ubiquitous. They should also spread evenly through the circuit, as in a *bus*, to avoid introducing unmatched series voltages that could otherwise emitter- or source-degenerate transistors unevenly. Although matching devices in this circuit is not critical, reliability and efficiency (because unpredictability leads to higher currents than necessary) require nominal matching performance from biasing transistors, current mirrors, and differential pairs (i.e., transistors with the same orientation, similar geometries, modular layout, and side-by-side placement).

6.3 Error Amplifier

Power efficiency and accuracy are vital parameters in the design of error amplifier A_{EA}. To start, as with power device S_O and buffer A_{BUF}, A_{EA} must survive low V_{IN} conditions (when the voltage across S_O is low) and demand low quiescent current I_Q for higher power efficiency and extended battery life. A_{EA} must also respond quickly (i.e., with

high bandwidth BW_{EA}) to dampen the adverse ac-accuracy effects of fast load dumps and supply ripple (as in power-supply rejection PSR) on regulator output v_O. What is more, dc-accuracy characteristics like load regulation (LDR), line regulation (LNR), and initial accuracy depend on reference V_{REF} and A_{EA}, which means A_{EA}'s open-loop gain must be high (but not high enough to compromise stability) and systematic and random input-referred offset V_{OS} low. Architecturally, A_{EA}'s output stage (and output voltage v_{EA}) must also consider the current- and voltage-drive demands of A_{BUF}. As a result, in summary, $V_{IN(min)}$ must be low, I_Q low, BW_{EA} high, PSR high (i.e., good), A_{EA} high, V_{OS} low, and v_{EA} compatible with the driving demands of A_{BUF}.

6.3.1 Low Headroom (i.e., Low Input Supply)

A_{EA}'s input common-mode range (ICMR) requirement, as set by V_{REF}, has an impact on the minimum input voltage (i.e., $V_{IN(min)}$) the circuit can accommodate without losing considerable gain. *Headroom* limit $V_{IN(min)}$ defines the point beyond which further decreases in V_{IN} push one or more high-gain transistors into their low-gain state: triode. More to the point, assuming the input stage is the limiting agent, V_{REF} and the minimum headroom allowed by A_{EA}'s upper ICMR point (e.g., V_{ROOM+} in Fig. 6.22*a*) determine A_{EA}'s $V_{IN(min)}$:

$$V_{IN(min)} = V_{REF} + V_{ROOM+} \tag{6.10}$$

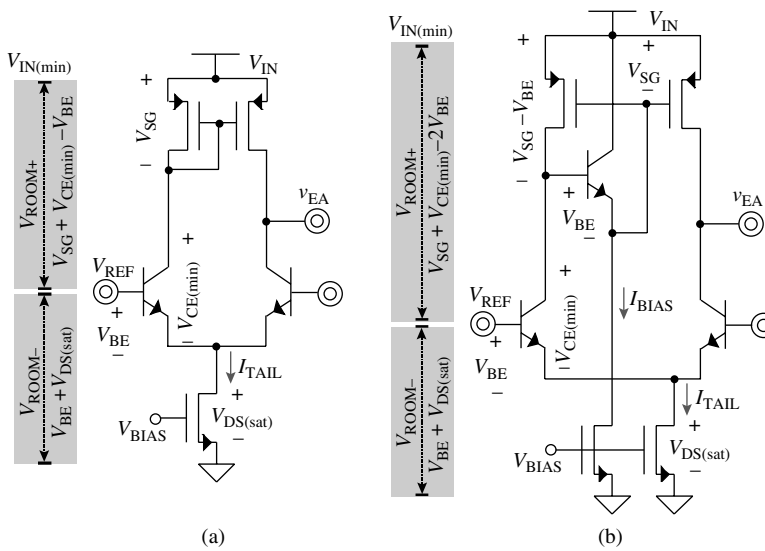


Figure 6.22 Headroom limits (i.e., $V_{IN(min)}$) for n-type input differential pairs loaded with (*a*) basic p-type and (*b*) complementary hybrid supply-referenced mirrors.

which is why low-voltage references synergize with low V_{IN} requirements.

Decreasing V_{REF} to minimize $V_{IN(min)}$ necessarily pushes A_{EA} toward one of its ICMR extremes. For example, to keep $V_{IN(min)}$ low, V_{REF} should only be high enough to bias an n-type differential pair just above the triode point of its single-transistor n-type tail-current sink, clearing negative headroom limit V_{ROOM-}:

$$V_{REF.NPN} \approx V_{ROOM-} \approx V_{BE(cold)} + V_{CE(min)} \approx 1\text{ V} \tag{6.11}$$

or

$$V_{REF.NMOS} \approx V_{ROOM-} \approx V_{GS(hot)} + V_{DS(sat)} = V_{TN} + 2V_{DS(sat)} \approx 1.25\text{ V} \tag{6.12}$$

where $V_{BE(cold)}$ and $V_{GS(hot)}$ correspond to the worst-case (i.e., highest) base-emitter and gate-source voltages, respectively. Similarly, V_{REF} should be low enough to bias a p-type differential pair just below the triode point of its single-transistor p-type tail-current source (with respect to $V_{IN(min)}$), clearing positive headroom limit V_{ROOM+}:

$$V_{REF.PNP} \le V_{IN(min)} - V_{ROOM+} = V_{IN(min)} - V_{EC(min)} - V_{EB(cold)} \approx V_{IN(min)} - 1\text{ V} \tag{6.13}$$

or

$$\begin{aligned} V_{REF.PMOS} &\le V_{IN(min)} - V_{ROOM+} \le V_{IN(min)} - V_{SD(sat)} - V_{SG} \\ &\approx V_{IN(min)} - 2V_{SD(sat)} - |V_{TP}| \approx V_{IN(min)} - 1.25\text{ V} \end{aligned} \tag{6.14}$$

Although not always the case, most references derive their outputs from the characteristic bandgap energy of silicon, which is why they normally generate voltages near 1.2 V. Attaching a resistor string from the bandgap point to ground with user-defined taps, however, divides this voltage to any desired value below 1.2 V, so using a tap point as V_{REF} (i.e., V_{REF} is below 1.2 V) increases A_{EA}'s flexibility with respect to ICMR and $V_{IN(min)}$. Similarly, increasing the value of V_{REF} is also possible by attaching the 1.2-V high-impedance bandgap point somewhere in the middle of the resistor string (between V_{REF} and ground) so the voltage divider can effectively amplify 1.2 V to the desired level, just as an operational amplifier would in a noninverting feedback configuration.

The only problem with low reference voltages is noise coupled through the common substrate and injected directly by other switching blocks in the system such as switching power supplies, data converters, digital signal-processing (DSP) circuits, and so on. The point is reducing V_{REF} does not decrease its noise content, which means

noise becomes a larger fraction of V_{REF} when V_{REF} is lower (i.e., *signal-to-noise ratio*, or SNR for short, is lower). As a result, because the linear regulator multiplies V_{REF} and all its noise to its target output voltage v_O by ratio V_O/V_{REF}, higher (i.e., better) SNR results when V_{REF} is high and ratio V_O/V_{REF} is low, which means, given equivalent $V_{IN(min)}$ limits, using an n-type differential pair whenever possible often produces better SNR results.

Low V_{REF} values alone, however, do not guarantee low $V_{IN(min)}$ limits, as not only positive headroom limit V_{ROOM+} plays a critical role in $V_{IN(min)}$ but so does the rest of A_{EA} (and the regulator, for that matter). Applying a basic p-type mirror load to an n-type differential input pair, for instance, as shown in Fig. 6.22*a*, places an emitter-base or, in this case, source-gate voltage V_{SG} in the ICMR path from V_{REF} to V_{IN}. The hybrid supply-referenced mirror load in Fig. 6.22*b* counters this V_{SG} by base-emitter voltage V_{BE} and consequently reduces the V_{SG} drop across the mirror load to $V_{SG} - V_{BE}$, except this technique only works if V_{SG} is greater than V_{BE} across the entire temperature range for all process corners, which is not always necessarily true. Replacing the level-shifting npn BJT with an n-type MOSFET whose threshold and saturation voltages V_{TN} and $V_{DS(sat)}$ combined exceed those of V_{SG} achieves similar objectives with a lower risk of pushing the mirroring PMOS into triode because p- and n-type FET parameters track better across process and temperature than FET parameters track BJT's.

A diode-connected BJT or MOSFET (i.e., collector-base or drain-gate shorted transistor) and a single-transistor current load constitute one of the most fundamental building blocks in an analog circuit so the lowest possible headroom limit a regulator can achieve is the sum of one V_{BE} or V_{GS} and one $V_{CE(min)}$ or $V_{DS(sat)}$. While V_{REF} (the minimum value of which negative headroom limit V_{ROOM-} sets) and positive headroom limit V_{ROOM+} in the basic configuration shown in Fig. 6.22*a* define a headroom limit (e.g., $V_{IN(min).a}$) that is worse than the fundamental extreme just mentioned:

$$V_{IN(min).a} \geq V_{REF} + V_{ROOM+} \geq V_{ROOM-} + V_{ROOM+} = V_{SG} + V_{DS(sat)} + V_{CE(min)} \tag{6.15}$$

the same is not necessarily true for the hybrid mirror case in Fig. 6.22*b* because V_{BE} ultimately offsets V_{SG} and decreases the resulting limit to

$$V_{IN(min).b} > V_{REF} + V_{ROOM+} \geq V_{ROOM-} + V_{ROOM+} = V_{SG} + V_{DS(sat)} + V_{CE(min)} - V_{BE} \tag{6.16}$$

In practice, however, the margin necessary to overcome the inherent mismatches in V_{SG} and V_{BE} with respect to process (e.g., how $|V_{TP}|$ and $V_{SD(sat)}$ match V_{BE} or how $|V_{TP}|$ and hole mobility μ_h match V_{TN} and electron mobility μ_e in the case of an NMOS level shifter) and

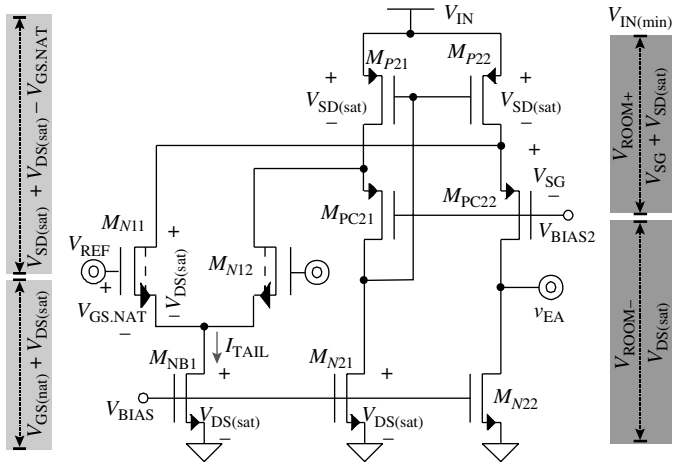


FIGURE 6.23 Headroom limit (i.e., $V_{IN(min)}$) for a natural NMOS differential pair loaded with a folded supply-referenced mirror.

temperature dependence typically increases $V_{IN(min).b}$ past the fundamental limit. Note $V_{IN(min)}$ is tightly coupled to V_{REF} in these two cases so variations in V_{REF} further increase (i.e., degrade) $V_{IN(min)}$.

Decoupling V_{REF} from the headroom path cannot only decrease $V_{IN(min)}$ but also relax the requirements on V_{REF}. The circuit shown in Fig. 6.23 achieves this by (1) folding the load to reduce positive headroom voltage V_{ROOM+} and (2) using a low-threshold device (e.g., natural NFET in this case) to lower negative headroom voltage V_{ROOM-}. Given the nature of the fundamental diode-connected limit described earlier, however, reducing $V_{IN(min)}$ to one V_{GS} and one $V_{DS(sat)}$ is difficult because every current path in the circuit experiences similar or worse constrains. Consider bias voltage V_{BIAS2} in Fig. 6.23 and its corresponding series path to V_{IN}, for example: the minimum headroom is the sum of current-source source-drain voltage $V_{SD.P22}$, cascode source-gate voltage $V_{SG.PC22}$, and the $V_{DS(sat)}$ that must exist between V_{BIAS2} and ground, the sum of which surpasses the fundamental limit of $V_{GS} + V_{DS(sat)}$. Note, however, cascode devices M_{PC21}–M_{PC22} need not match well (as will be discussed in Sec. 6.2.3) so their aspect ratios can be relatively high and their corresponding saturation voltages low.

The current-source path to ground through M_{P21} and M_{N21} and the input stage also present similar $V_{IN(min)}$ limits, considering natural NFETs M_{N11} and M_{N12} mitigate the impact of V_{GS} in the latter. While M_{P21}'s V_{GS} and M_{N21}'s $V_{DS(sat)}$ present the fundamental limit, the input stage path does not because, as with V_{BIAS2}, the reference circuit necessarily includes at least one $V_{EC(min)}$ or $V_{SD(sat)}$ voltage between V_{IN} and V_{REF} to ensure V_{REF} remains constant with respect to ground. As a result, A_{EA}'s input path,

which consists of $V_{GS,N11}$ and $V_{DS,B1}$, also includes another $V_{SD(sat)}$ from the reference circuit. Natural NFETs in the differential pair ultimately reduce the V_{GS} component by nearly one V_{TN}, negating the adverse effects of the additional $V_{DS(sat)}$ and supplying margin, in return, for lower aspect ratios and, in consequence, better matching performance in M_{N11}–M_{N12} (i.e., longer channel lengths decrease mismatches). In summary, n-type input stages tend to increase (i.e., improve) SNR, natural NFETs relax the requirements on V_{REF}, and folding architectures decrease (i.e., improve) headroom (i.e., $V_{IN(min)}$) limits.

6.3.2 High Power-Supply Rejection

N-Type Power Device

Power-supply rejection (PSR) superimposes architectural requirements on the loading mirror of error amplifier A_{EA} that primarily depend on the power pass device the regulator uses. Because n-type pass transistors are in a voltage-follower configuration, for instance, and they impress whatever ac signal is present at their inputs onto v_O, A_{EA} must suppress as much of the supply ripple as possible in v_{EA}. Ground-referenced mirror loads, as discussed in the previous chapter and shown in Fig. 6.24, do just this, subtract the supply ripple injected into v_{EA}. To be more explicit, M_{M2} transistors in Fig. 6.24*a*, *b*, and *c* subtract the supply ripple injected from v_{IN} to v_{EA} by M_{Tail}–M_{D2} in Fig. 6.24*a*, M_{B2} in Fig. 6.24*b*, and M_{B2}–M_{C2} in Fig. 6.24*c* because M_{M2} devices receive similar ripples at their respective inputs. As a result, to first order, ground-referenced mirror loads produce supply gains (i.e., A_{IN} or v_{EA}/v_{IN}) that near 0 V/V and PSRs that tend to infinity.

P-Type Power Device

As veritable complements, the input of a p-type power device must receive the same supply ripple already present at its emitter or source to keep the ripples common mode (i.e., in phase) with respect to emitter-base or source-gate terminals, in other words, to prevent the transconductance of the device from amplifying any difference in ripples. Much like ground-referenced mirrors subtract supply ripple, the supply-referenced counterparts shown in Fig. 6.25 and discussed in the previous chapter source the fraction necessary to fully impress v_{IN}'s ripple onto v_{EA}. In more specific terms, the transconductances in M_{M2} devices in Fig. 6.25*a*, *b*, and *c*, which constitute the mirror portions of the transistors, introduce approximately half the ripple present in v_{IN} into v_{EA} (because their inputs source that much equivalent current) while the output resistances (i.e., r_{ds}) in the supply path inject the other half. Ultimately, supply-referenced mirror loads yield supply gains (i.e., A_{IN} or v_{EA}/v_{IN}) and PSRs near 1 V/V and carry ac supply ripple v_{in} in v_{EA}.

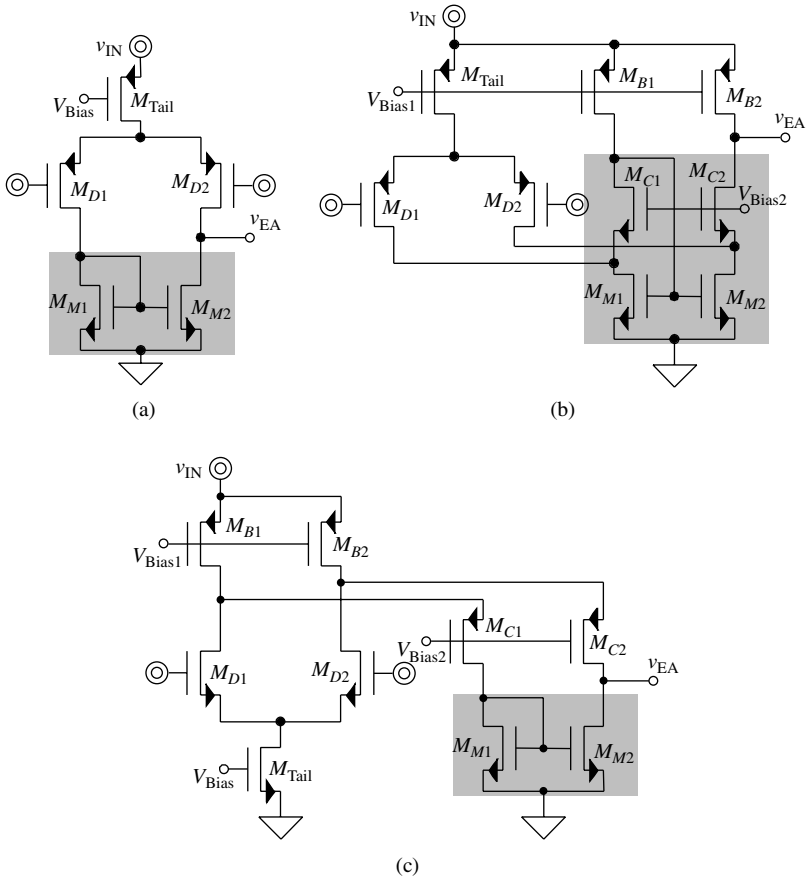


FIGURE 6.24 (*a*) and (*b*) P-type and (*c*) n-type differential input pairs with (*a*) and (*c*) basic and (*b*) folding mirror loads referenced to ground, which are suitable for n-type power pass devices.

Generalities

The canceling and summing effects of these ground- and supply-referenced mirrors hinges on the symmetry of the circuit and the mirroring features of the load, so keeping the design symmetrical and the mirror as ideal as possible are important. The overriding assumption in the previous two subsections, however, is that the input resistance of the mirrors (e.g., $1/g_m$) is substantially lower than the resistances driving them because the latter resistances redirect and distort the current the mirror is able to reproduce with respect to the supply-injected current already present at the output. As such, PSR generally improves when the impedances to supply rails v_{IN} and ground are high, which means high Early-voltage BJTs and long channel-length MOSFETs favor high PSR performance.

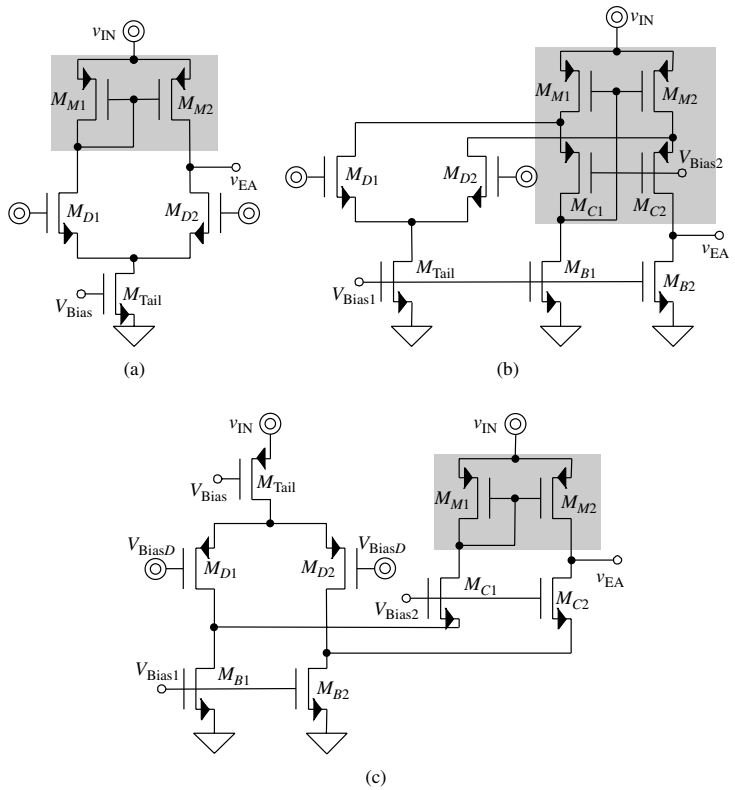


FIGURE 6.25 (*a*) and (*b*) N-type and (*c*) p-type differential input pairs with (*a*) and (*c*) basic and (*b*) folding mirror loads referenced to supply v_{IN}, which are suitable for p-type power pass devices.

6.3.3 Low Input-Referred Offset

Any input-referred dc offset present in A_{EA} (i.e., V_{OS}) translates directly to a dc offset in v_O because the noninverting negative-feedback configuration of the linear regulator (Fig. 6.1) amplifies V_{OS} (along with reference V_{REF}) by user-defined or application-specific ratio V_O/V_{REF}, which means V_{OS} degrades regulator accuracy. As such, keeping V_{OS} low is imperative, especially when considering its dependence to temperature compounds its adverse effects on v_O's accuracy. However, while keeping the systematic component in V_{OS} low is mostly a circuit-based exercise, random variations in V_{OS} depend on both the circuit and its silicon-based embodiment.

Systematic

The most common cause of systematic offset in analog circuits is asymmetry in the circuit. All the amplifiers in Figs. 6.24 and 6.25, for

instance, would be perfectly symmetrical if their respective outputs (i.e., v_{EA} terminals) were to duplicate the dc voltages already present on the other side of their load mirrors, which normally falls within one base-emitter or gate-source voltage from either supply. Otherwise, if a voltage mismatch ΔV exists, the gain across the error amplifier attenuates ΔV by its gain (i.e., A_{EA}) and introduces a systematic input-referred offset $V_{OS.S}$ at its input that is equivalent to

$$V_{OS.S} = \frac{\Delta V}{A_{EA}} \tag{6.17}$$

which is why keeping A_{EA} as high as possible, given the stability constraints of the system, is important. For example, a 0.5 V mismatch at the output of a 40-dB (i.e., 100-V/V) amplifier introduces a 5-mV systematic offset at the input of the amplifier, which degrades the room-temperature accuracy of a 1.2 V reference by roughly half a percent (if not trimmed) and the temperature-drift performance by $V_{OS.S}$'s drift across temperature.

Reducing the impact of $V_{OS.S}$ on accuracy would be easier if the gain across A_{EA} had not been constrained by the stability requirements of the regulator and dc offsets in v_{EA} had not been inherent. Because A_{EA} in n-type power-transistor applications is normally higher than in their p-type counterparts (given n-type followers contribute little to no voltage gain and A_{EA} must therefore amplify the deficit), regulators with n-type power devices tolerate higher systematic offset voltages in v_{EA}. For example, an NMOS ground-referenced mirror dictates V_{EA} should be a V_{GS} above ground (e.g., 1 V), yet a 2.5-V output with an npn power device sets V_{BUF} at roughly 3.2 V and a PMOS follower buffer translates that to a V_{EA} of 2.2 V, resulting in a 1.2-V offset in v_{EA} and a 12-mV systematic input-referred offset (assuming A_{EA} is 40 dB). Similarly, a PMOS supply-referenced mirror dictates V_{EA} should be one V_{EB} or V_{SG} below V_{IN}, and although a p-type power transistor sets V_{BUF} at one V_{EB} or V_{SG} below V_{IN}, A_{BUF} introduces an offset of 0.3–1.2 V, which when divided by 40 dB, for example, translates to 3–12 mV of systematic input-referred offset.

Wide dc voltage swings at the input of power device S_O (i.e., at v_{BUF} and consequently at v_{EA}) in response to large output-current variations exacerbate the problem. Assuming the regulator remains out of dropout (i.e., in the linear region), A_{EA} is 40 dB, and considering overdrive voltages of 250–500 mV in S_O in response to 50–100-mA load currents are typical, systematic input-referred offset voltage variations of 2.5–5 mV are not uncommon. This offset increases as the regulator approaches the dropout region, where gain drops drastically and $V_{OS.S}$ increases proportionately. Note this load-dependent variation in $V_{OS.S}$ (i.e., $\Delta V_{OS.S}$) represents a load-regulation

(LDR) effect because changes in I_O lead to steady-state fluctuations in Δv_{BUF} and ultimately in v_O (via $V_{OS.S}$), so a more accurate depiction of LDR is

$$\begin{aligned}
\text{LDR} \equiv \frac{\Delta v_O}{\Delta I_O} &= \Delta I_O R_{O.CL} + \Delta V_{OS.S}\left(\frac{V_O}{V_{REF}}\right) \\
&= \Delta I_O R_{O.CL} + \left(\frac{\Delta v_{EA(max)}}{A_{EA}}\right)\left(\frac{V_O}{V_{REF}}\right) \\
&\approx \Delta I_O R_{O.CL} + \left(\frac{\Delta v_{SG.PO(max)}}{A_{EA}}\right)\left(\frac{V_O}{V_{REF}}\right) \\
&\approx \Delta I_O R_{O.CL} + \left[\left(\frac{1}{A_{EA}}\right)\left(\frac{\Delta I_{O(max)}}{g_{m.O(avg)}}\right)\right]\left(\frac{V_O}{V_{REF}}\right)
\end{aligned} \tag{6.18}$$

where $\Delta V_{OS.S}$ degrades LDR, $R_{O.CL}$ refers to the closed-loop output resistance of the regulator, $g_{m\cdot O(avg)}$ is S_O's average transconductance (e.g., evaluated at $0.5I_{O(max)}$), and $\Delta I_{O(max)}$ represents the worst possible full-scale dc variation in i_O, which produces $\Delta v_{SG.PO(max)}$ (i.e., $\Delta v_{BUF(max)}$ or equivalently, if A_{BUF} is 1 V/V, $\Delta v_{EA(max)}$).

A way to mitigate the swinging effects of v_{BUF} on v_{EA} in response to large-signal variations in output current i_O, in other words, to reduce dv_{EA}/di_O and its resulting LDR degradation is to shift v_{BUF} by an oppositely phased load-dependant voltage drop. The dynamically adaptive positive-feedback n-type follower buffers for PMOS power device M_{PO} shown in Figs. 6.19 and 6.20 do just this. To start, as with all power p-type devices, the loop and the pass transistor react to an increase in I_O by decreasing the dc level of v_{BUF} so dv_{BUF}/di_O is negative. In response to this drop, the current sensor (i.e., M_{PS}) and mirror (i.e., M_{N1}–M_{N2}) combination that comprise the positive-feedback loop in buffer A_{BUF} pull more drain current $I_{D.NBUF}$ from follower M_{NBUF}, increasing its corresponding gate-source voltage $V_{GS.NBUF}$ and producing a positive $dv_{GS.NBUF}/di_O$. The result is positive dc changes in M_{NBUF}'s v_{GS} compensate corresponding negative dc variations in v_{BUF}, reducing the overall large-signal variability of v_{EA} with respect to I_O and its effect on $V_{OS.S}$ and LDR:

$$\begin{aligned}
\frac{dv_{EA}}{di_O} &= \frac{dv_{GS.NBUF}}{di_O} + \frac{dv_{BUF}}{di_O} = \frac{dv_{GS.NBUF}}{di_O} - \frac{dv_{SG.O}}{di_O} \\
&\approx \left(\frac{1}{g_{m.NBUF(avg)}}\right) - \left(\frac{1}{g_{m.PO(avg)}}\right)
\end{aligned} \tag{6.19}$$

where $g_{m.\mathrm{NBUF(avg)}}$ and $g_{m.\mathrm{PO(avg)}}$ are M_{NBUF}'s and M_{PO}'s average transconductances across I_O's entire range. Interestingly, over-compensating the circuit by allowing $|dv_{\mathrm{GS.NBUF}}/di_O|$ to exceed $|dv_{\mathrm{SG.}O}/di_O|$ (assuming the change is benign to the rest of the circuit) improves LDR performance (i.e., dc accuracy) because a negative systematic offset, which amounts to a dc load-dependent voltage shift across A_{BUF} that produces a correspondingly higher Δv_{EA} during heavier loads, offsets the low loop-gain effect $\Delta I_O R_{O.\mathrm{CL}}$ produces:

$$\begin{aligned}\mathrm{LDR} \equiv \frac{\Delta v_O}{\Delta I_O} &= \Delta I_O R_{O.\mathrm{CL}} + \left(\frac{\Delta v_{\mathrm{EA}}}{A_{\mathrm{EA}}}\right)\left(\frac{V_O}{V_{\mathrm{REF}}}\right) = \Delta I_O R_{O.\mathrm{CL}} \\ &+ \left(\frac{\Delta v_{\mathrm{BUF(max)}} - \Delta v_{\mathrm{EA(max)}}}{A_{\mathrm{EA}}}\right)\left(\frac{V_O}{V_{\mathrm{REF}}}\right) \approx \Delta I_O R_{O.\mathrm{CL}} \\ &+ \frac{1}{A_{\mathrm{EA}}}\left(\frac{\Delta I_O}{g_{m.\mathrm{PO(avg)}}} - \frac{\Delta I_O}{K_A g_{m.\mathrm{NBUF(avg)}}}\right)\left(\frac{V_O}{V_{\mathrm{REF}}}\right) \end{aligned} \tag{6.20}$$

where M_{NBUF} carries a fraction of I_O that is equivalent to a mirror attenuation between M_{PO} and M_{PS} of K_A^{-1}.

Random

Random input-referred offset $V_{\mathrm{OS}.R}$, like its systematic counterpart, results from asymmetry in the circuit, but with respect to process- and temperature-induced mismatches of otherwise identical components instead of circuit-based differences. The problem is mobility, oxide thicknesses, doping concentrations, threshold voltages, and other process-dependent parameters not only differ from lot to lot and die to die but also from transistor to transistor across a single die. The fact is temperature and diffusion gradients and dissimilar silicon-profile conditions across the die cause transistor parameters to differ. These mismatches may be relatively smaller when constrained to smaller areas but their effects are nonetheless apparent in high-precision regulators, where only a 6–12 mV of input-referred offset produces a 0.5–1% degradation in regulator accuracy.

The first approach in battling $V_{\mathrm{OS}.R}$ is to reduce physical offset, in other words, to improve the matching performance of critical devices in the circuit. In determining which devices are critical, viewing a current offset ΔI between two supposedly matched transistors as the result of an equivalent input-referred offset voltage v^* at one of the bases or gates (i.e., v^* is $\Delta I/g_m$) helps. As in noise analysis, translating the effects of all offset voltages forward to the output (e.g., to v_{EA} or i_{EA}) and referring the combined result back to the input yields random

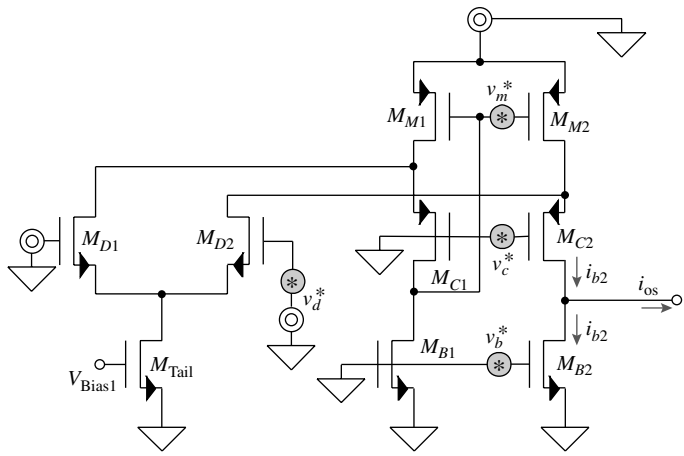


Figure 6.26 Error amplifier A_{EA}'s random-mismatch small-signal equivalent circuit.

offset voltage $V_{OS.R}$, the input-referred offset effects of all random mismatches. These random variations appear between every matched pair of transistors in the circuit, as shown in Fig. 6.26, and are small enough in magnitude to constitute small-signal (ac) perturbations. As a reference, statistically meaningful [i.e., three-sigma (3σ)] values for v^* in well matched BJTs and MOSFETs are typically 1–3 mV and 5–15 mV, respectively, which means BJTs tend to match better than MOSFETs.

From an ac-signal viewpoint, when considering offsets in the error amplifier (i.e., A_{EA}) shown in Fig. 6.26, input pair M_{D1}–M_{D2}, mirror M_{M1}–M_{M2}, and bias devices M_{B1}–M_{B2} constitute common-source amplifiers to their respective offsets so they produce similar effects on the overall offset current that results (i.e., i_{os}), which would otherwise be zero in the case of perfectly matched devices. Since mirror transistors M_{M1}–M_{M2} source-degenerate cascode transistors M_{C1}–M_{C2}, an input-referred offset voltage in the cascodes has a lower (i.e., source degenerated) and often negligible impact on i_{os}, when compared to the influence exerted by other transistors in the circuit. When combined, because random uncorrelated offsets are probabilistic in nature, their effects do not sum linearly but as the square-root sum of the squares, which is why keeping i_{os} and its constituent terms squared (e.g., i_{os}^2) is a popular practice:

$$i_{os}^2 \approx v_d^{*2} g_{m.D}^2 + v_m^{*2} g_{m.M}^2 + v_b^{*2} g_{m.B}^2 + \frac{v_c^{*2} g_{m.C}^2}{(1 + g_{m.C} r_{ds.M})^2}$$

$$\approx v_d^{*2} g_{m.D}^2 + v_m^{*2} g_{m.M}^2 + v_b^{*2} g_{m.B}^2 \qquad (6.21)$$

Dividing i_{os} (or i_{os}^2) by input-pair transconductance $g_{m.D}$ (or $g_{m.D}^2$) refers (i.e., translates) the resulting offset current to the input as $V_{OS.R}$:

$$V_{OS.R} = \frac{i_{os}}{g_{m.D}} \approx \sqrt{\frac{v_d^{*2} g_{m.D}^2 + v_m^{*2} g_{m.M}^2 + v_b^{*2} g_{m.B}^2}{g_{m.D}^2}}$$

$$= \sqrt{v_d^{*2} + v_m^{*2}\left(\frac{g_{m.M}^2}{g_{m.D}^2}\right) + v_b^{*2}\left(\frac{g_{m.B}^2}{g_{m.D}^2}\right)} \tag{6.22}$$

Generally, given their impact on $V_{OS.R}$, input pairs M_{D1}–M_{D2}, load mirrors M_{M1}–M_{M2}, and in the case of folded architectures, biasing transistors M_{B1}–M_{B2} in Figs. 6.24 and 6.25 must match well for $V_{OS.R}$ to remain low.

Because transistor transconductance ratios, as observed in $V_{OS.R}$'s expression, transpose individual offsets back to the input, the second approach to reducing $V_{OS.R}$ is circuit based, by carefully designing all g_m's (i.e., by choosing optimal bias currents and transistor aspect ratios). For instance, increasing the differential pair's transconductance (i.e., $g_{m.D}$) and decreasing both the mirror and biasing transistors' respective transconductances (i.e., decreasing $g_{m.M}$ and $g_{m.B}$) mitigate the detrimental effects offsets in the mirror and biasing transistors have on $V_{OS.R}$. Unfortunately, there is no straightforward circuit-design means of decreasing the impact of an offset in the input pair because v_d^* is already present at A_{EA}'s input. The only way of reducing this offset is through costly offset-calibration schemes such as dynamic-element matching, chopper techniques, trimming, and others, which add complexity, power losses, switching noise, test time, and/or silicon real estate. This is not to say, however, designers do not employ such offset-cancellation schemes, because they do, as in, for instance, many analog-to-digital (A/D) converters, but energy-constrained applications, unfortunately, leave little power margin for linear regulators to adopt them. Alternatively, shifting the trimming point from V_{REF} to regulator output v_O (or equivalently, to feedback node v_{FB}) trims (i.e., cancels) the combined effects of random variations in V_{REF} and A_{EA}. The problem with this latter approach is V_{REF} is no longer trimmed on its own, that is to say, V_{REF} alone is less accurate, which means the rest of the system must now conform by either sacrificing accuracy performance or requiring more power and silicon area to implement a separate (more accurate) reference.

6.3.4 Layout

Matching Hierarchy

Perhaps the most important layout notes for error amplifier A_{EA} relate to random mismatches. As such, categorizing and grouping transistors

with respect to their matching requirements is the first step in the layout process. In this respect, the highest (i.e., first) matching degree applies to "critical" devices, which should conform, if and when possible, to the following seemingly extreme guidelines:

1. Place transistors close to one another to minimize the total variation spread resulting from two-dimensional gradients on the die.
2. Build transistor array as modular (i.e., square) as possible to reduce the silicon area and minimize the parameter spreads resulting from two-dimensional gradients on the die.
3. Orient transistors in the same direction to ensure fabrication effects (from deposition, etc.) are consistent among matching transistors.
4. Employ a common-centroid layout strategy (i.e., define a point in the layout to be the common center of "mass" for matching devices) to balance the effects of two-dimensional gradients.
5. Cross-couple equally sized transistors to cancel the effects of two-dimensional gradients on the die.
6. Place small dummy transistors around the peripheral transistors in the array at a distance that is equivalent to the intra-transistor distance within the array to mitigate proximity-mismatch effects.
7. Avoid placing metal routes (i.e., lines) immediately above matching transistors to prevent packaged-induced stress fields from creating localized and mismatched piezoelectric effects. If routing above the transistors is unavoidable, match the routing tracks so that all matched transistors experience similar effects.
8. Place one or two uniform sheets of one or two top-level metals immediately above the entire array to spread the effects of packaged-induced localized stress fields equally.

Deciding which subset and when and how to apply the aforementioned measures is subjective and dependent on the target specifications demanded and the technology used to fabricate the circuit. Nevertheless, the second highest matching degree usually backs off from the extreme measures that often demand considerably more silicon real estate, as in, for example, measures (4) to (8) in the list above, while retaining "good" matching practices, like maybe those specified in (1) to (3). The last (or third) matching degree refers to "nominal" matching devices, in which case (1) may be sufficient.

In contemplating the matching requirements of error amplifier A_{EA}, input pair M_{D1}–M_{D2}, load mirror M_{M1}–M_{M2}, and folding bias

transistors M_{B1}–M_{B2} in Figs. 6.24 and 6.25 must all match as well as possible so "critical" matching considerations apply. Although the matching performance of tail current source M_{Tail} to folding bias transistors M_{B1}–M_{B2} in Figs. 6.24*b* and 6.25*b* does not affect random offset performance, random mismatches between M_{Tail} and bias pair M_{B1}–M_{B2} produce absolute variations in quiescent current, unity-gain frequency, slew-rate current, and others, which ultimately increase performance spreads and decrease die yield. (Die yield refers to the fraction of dies in a wafer that meet target specifications, which translates to revenue per wafer.) M_{Tail} should therefore match bias pair M_{B1}–M_{B2}, but not to the same degree critically matched transistors do, so "good" matching practices normally suffice. The matching performance of cascode devices M_{C1}–M_{C2} may exert little influence on random and systematic offsets and other circuit-performance parameters, but completely neglecting them may increase their impact to noticeable levels, so applying "nominal" matching practices is often prudent.

Notes

Other important layout notes relate to bandwidth and noise, and to a lesser extent, power and thermal effects. To start, introducing parasitic capacitance to A_{EA}'s output, because of its high output resistance, slows the circuit so all transistors connected to v_{EA} should be as close as possible, which means A_{EA} and A_{BUF} should be close or integrated into one layout. Similarly, and to avoid noise injection from the substrate and other electrical lines in the system, A_{EA}'s inputs should be as close as possible to their respective bond pads. In fact, the sense path from regulator output v_O, which ideally sits at the point of load (PoL), to A_{EA}'s input should be as short as possible and surrounded by low-impedance tracks to shunt and steer incoming noise away from A_{EA}.

Series voltage drops in the supplies introduce systematic offset voltages in supposedly matched transistors so power-supply buses of maybe 5–8 μm in width should surround A_{EA}. In traditional operational amplifier designs, displacing differential pairs from power output devices is normally a good engineering practice to mitigate the impact of the thermal gradients "hot spots" in the die induce. However, the pass transistor in a linear regulator normally constitutes about 70% of the layout, which means displacing the differential pair is not practical, especially considering modern plastic packages have low thermal impedances so the temperature gradients they induce are substantially low.

Design Example 6.3 Refer to the specifications and process parameters outlined in Table 6.4 to design a 6-μA error amplifier for the dynamically adaptive positive-feedback buffer and PMOS power transistor designed in Examples 6.1 and 6.2. The complete internally compensated low-dropout (LDO) regulator

Circuit Parameter	Specification	PMOS Process Parameter	Value
V_{IN}	1.1–1.6 V	$\lvert V_{TP}\rvert$ and V_{TN}	0.6 V ± 150 mV
V_O	1 V	$V_{TN.NAT}$	0 V ± 150 mV
V_{REF}	0.9 V	K'_P	40 μA/V² ± 20%
A_{EA}	≈ 40 dB	K'_N	100 μA/V² ± 20%
I_Q	≤ 6 μA	L	≥ 0.35 μm
$V_{EA(max)}$	V_{IN} – 0.3 V	C''_{OX}	15 fF/μm²
$V_{EA(min)}$	0.2 V	$\beta_{L(min)}$	500 mV^{-1}
f_{0dB}	1 MHz	$\lambda_{3L(min)}$	10 mV^{-1}
$V_{OS.R}$	≤ 25 mV	$\beta_{0.PNP}$	50–150 A/A
$V_{OS.S}$	≤ 10 mV	$V_{A.PNP}$	15 V

TABLE 6.4 Target Specifications and Process-Parameter Values for the Error Amplifier in Example 6.3

should generate a 1-V output with a unity-gain frequency of 1 MHz from a 0.9-V reference and a 0.9–1.6-V NiMH battery, the practical operating range of which is 1.1–1.6 V. The output of the amplifier should swing from 0.2 V to within 0.3 V of V_{IN} and incur no more than 25 mV of random uncorrelated input-referred offset and 10 mV of systematic offset.

Architecture notes

1. Input differential pair:
   a. Since V_{REF} is 0.9 V, use natural NMOSFETs to relax the requirements imposed on the tail current-sink device.
   b. Since $V_{TN.NAT(min)}$ is –0.15 V, connect bulk and ground terminals together to allow bulk effects to increase $V_{TN.NAT}$ to a positive value under worst-case conditions so that the source-coupled node remains below V_{REF} and the differential pair remains saturated (i.e., out of triode).
2. Load
   a. Since the power device is a PMOSFET, use a supply-referenced load mirror to apply a common-mode supply ripple to the PMOSFET's gate for higher PSR.
   b. Since $V_{IN(min)}$ is 1.1 V, use a folding mirror load, as shown in Fig. 6.27.
   c. Use a diode-connected PMOS transistor (i.e., M_{CB}) to bias the cascodes in the folding mirror.
3. Regulator output (given V_{REF} is 0.9 V and V_O should be 1 V):
   a. Use resistor divider R_{FB1}-R_{FB2} to set closed-loop gain V_O/V_{REF}:

$$\frac{V_O}{V_{REF}} = \frac{R_{FB1} + R_{FB2}}{R_{FB2}} \equiv \frac{(1)}{(0.9)} = 1.11$$

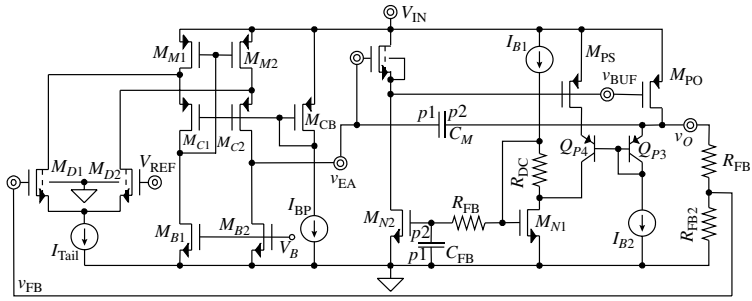


FIGURE 6.27 Error amplifier design for Example 6.3 and accompanying buffer and power PMOS transistor from Examples 6.1 and 6.2.

b. Set resistor bias current to 3 µA to ensure resistor string sinks M_{PO}'s subthreshold current during worst-case conditions and extreme process and temperature corners:

$$I_{R_O} = \frac{V_O}{R_{FB1} + R_{FB2}} \approx 3\,\mu A$$

$$\therefore \quad R_{FB2} \equiv 300\ k\Omega \text{ and } R_{FB1} \equiv 33.3\ k\Omega$$

4. Frequency compensation:
Set dominant pole at v_{EA} with Miller capacitor C_M (across A_{BUF} and M_{PO}).

5. Negative feedback:
Since M_{PO} presents an inverting gain stage, connect sensing feedback signal v_{FB} to A_{EA}'s noninverting input (i.e., gate of M_{D1}).

Transistor design

6. Low random offset:
   a. Use large differential pair and large mirror load (i.e., long channel lengths: $10L_{(min)}$):

   $\therefore \quad L_{D1} \equiv L_{D2} \equiv L_{M1} \equiv L_{M2} \equiv 10L_{(min)} \equiv 3.5\ \mu m$ and match critically well.

   b. Match biasing current sinks M_{B1}–M_{B2} well but design them to produce a small small-signal gain in A_{EA} (i.e., have them present low r_{ds}'s) to ensure the loop's unity-gain frequency f_{0dB} drops to 0 dB before encountering the parasitic poles of the regulator (for stability). In other words, match M_{B1}–M_{B2} well but use short channel lengths to induce higher channel-length modulation effects:

   $\therefore \quad L_{B1} \equiv L_{B2} \equiv L_{(min)} \equiv 0.35\ \mu m$ and match critically well.

7. Open-loop gain A_{EA}:

$$A_{EA.DC} \approx g_{m.D1} r_{ds.B2} \approx \frac{\sqrt{2I_{D1}K_N'\left(\frac{W}{L}\right)_{D1}}}{I_{B2}\lambda_{L(min)}} = \frac{\sqrt{2\left(\frac{I_{Tail}}{2}\right)K_N'\left(\frac{W}{L}\right)_{D1}}}{I_{B2}\lambda_{L(min)}} \equiv 100$$

$$\text{or} \quad \frac{\sqrt{I_{Tail}\left(\frac{W}{L}\right)_{D1}}}{I_{B2}} \equiv \frac{A_{EA.DC}\lambda_{L(min)}}{\sqrt{K_N'}} = \frac{(100)(0.1)}{\sqrt{(100\,\mu)}} \approx 1000$$

so choosing I_{Tail} as 2 μA and I_{B2} as 1 μA yields

$$\therefore \quad \left(\frac{W}{L}\right)_{D1} \approx \frac{(1000)^2(1\,\mu)^2}{(2\,\mu)} = 0.5 \qquad \text{or} \qquad \left(\frac{W}{L}\right)_{D1} \equiv \left(\frac{W}{L}\right)_{D2} \equiv \frac{1.75\,\mu m}{3.5\,\mu m}$$

8. $V_{EA(max)}$:

a. $V_{EA(max)} = V_{IN} - V_{SD.M2} - V_{SD.C2(sat)} \geq V_{IN} - 0.3\ V$

$$\therefore \quad V_{SD.M2(sat)} \equiv V_{SD.C2(sat)} \leq 0.15\ V$$

b. $$V_{SD.M2(sat)} = \sqrt{\frac{2I_{M2}}{K'_P\left(\frac{W}{L}\right)_{M2}}} \leq \sqrt{\frac{2(I_{D1}+I_{B2})}{K'_{P(min)}\left(\frac{W}{L}\right)_{M2}}} \leq 0.15\ V$$

$$\text{or} \qquad \left(\frac{W}{L}\right)_{M2} \geq \frac{2(I_{D1}+I_{B2})}{V^2_{SD.M2(sat)}K'_{P(min)}} = \frac{2(1\,\mu+1\,\mu)}{(0.15)^2(32\,\mu)} = 5.55$$

$$\therefore \quad \left(\frac{W}{L}\right)_{M1} \equiv \left(\frac{W}{L}\right)_{M2} \equiv 6 \equiv \frac{21\,\mu m}{3.5\,\mu m}$$

c. $$V_{SD.C2(sat)} = \sqrt{\frac{2I_{C2}}{K'_P\left(\frac{W}{L}\right)_{C2}}} \leq \sqrt{\frac{2I_{B2}}{K'_{P(min)}\left(\frac{W}{L}\right)_{C2}}} \leq 0.15\ V$$

$$\text{or} \qquad \left(\frac{W}{L}\right)_{C2} \geq \frac{2I_{B2}}{V^2_{SD.C2(sat)}K'_{P(min)}} = \frac{2(1\,\mu)}{(0.15)^2(32\,\mu)} = 2.8$$

$$\therefore \quad \left(\frac{W}{L}\right)_{C1} \equiv \left(\frac{W}{L}\right)_{C2} \equiv 3 \equiv \frac{1.05\,\mu m}{0.35\,\mu m}$$

d. Choose $V_{SD.M2}$ as 0.15 V:

$$V_{IN} - V_{SD.M2} - V_{SG.C2} = V_{IN} - V_{SG.CB}$$

$$\text{or} \qquad V_{SG.CB} = |V_{TP}| + \sqrt{\frac{2I_{BP}}{K'_P\left(\frac{W}{L}\right)_{CB}}} \equiv V_{SD.M2} + V_{SG.C2}$$

$$= V_{SD.M2} + |V_{TP}| + \sqrt{\frac{2I_{C2}}{K'_P\left(\frac{W}{L}\right)_{C2}}}$$

or choosing I_{BP} as 0.5 µA:

$$\left(\frac{W}{L}\right)_{CB} = \frac{2I_{BP}}{K'_P\left[V_{SD.M2} + \sqrt{\dfrac{2I_{C2}}{K'_P\left(\dfrac{W}{L}\right)_{C2}}}\right]^2} = \frac{2(0.5\,\mu)}{(40\,\mu)\left[0.15 + \sqrt{\dfrac{2(1\,\mu)}{(40\,\mu)\left(\dfrac{1.05}{0.35}\right)}}\right]^2} = 0.32$$

$$\therefore \quad \left(\frac{W}{L}\right)_{CB} \equiv \frac{1\,\mu m}{3\,\mu m}$$

9. $V_{EA(min)}$:

$$V_{EA(min)} = V_{DS.B2(sat)} \le \sqrt{\frac{2I_{B2}}{K'_{N(min)}\left(\dfrac{W}{L}\right)_{B2}}} \le 0.2\ V$$

or $$\left(\frac{W}{L}\right)_{B2} \ge \frac{2I_{B2}}{K'_{N(min)}V^2_{EA(min)}} = \frac{2(1\,\mu)}{(80\,\mu)(0.2)^2} = 0.62 \quad \therefore \quad \left(\frac{W}{L}\right)_{B2} \equiv \frac{2\,\mu m}{0.35\,\mu m}$$

10. f_{0dB} (which is roughly equal to gain-bandwidth product GBW):
Using power PMOS gain A_{PO} and A_{EA}'s output resistance R_{EA}, which is roughly $r_{ds.B2}$:

$$2\pi f_{0dB} \approx GBW = A_{EA}A_{BUF}A_{PO}(2\pi f_{P.EA}) \approx \frac{A_{EA}A_{BUF}A_{PO}}{C_M(A_{BUF}A_{PO})R_{EA}}$$

$$\approx \frac{(g_{m.D1}R_{EA})A_{BUF}A_{PO}}{C_M(A_{BUF}A_{PO})R_{EA}} = \frac{g_{m.D1}}{C_M} \equiv 2\pi(1\ MHz)$$

or $$C_M = \frac{g_{m.D1}}{2\pi f_{0dB}} = \frac{\sqrt{2I_{D1}K'_N\left(\dfrac{W}{L}\right)_{D1}}}{2\pi f_{0dB}} = \frac{\sqrt{2(1\,\mu)(100\,\mu)\left(\dfrac{1.75}{3.5}\right)}}{2\pi(1\ MHz)} = 1.59\ pF$$

$$\therefore \quad C_M \equiv 1.6\ pF$$

Design checks

11. I_Q:

$$I_Q = I_{M1} + I_{M2} + I_{BP} = (2\ \mu A) + (2\ \mu A) + (0.5\ \mu A) = 4.5\ \mu A,$$

which with a 30% variation, is 5.85 µA.

12. $V_{IN(min)}$ (assuming minimum voltage across I_{BP} is 0.2 V):

$$V_{IN(min)} = \text{MAX}\begin{cases} V_{SG.M1(max)} + V_{DS.B1(sat)} = |V_{TP(max)}| + V_{SD.M1(sat.max)} + V_{DS.B1(sat.max)} \\ V_{SG.CB(max)} + V_{I_{BP}} = |V_{TP(max)}| + V_{SD.CB(sat.max)} + 0.2\text{ V} \end{cases}$$

$$\approx \text{MAX}\begin{cases} |V_{TP(max)}| + \sqrt{\dfrac{2I_{M1}}{K'_{P(min)}\left(\dfrac{W}{L}\right)_{M1}}} + \sqrt{\dfrac{2I_{B1}}{K'_{N(min)}\left(\dfrac{W}{L}\right)_{B1}}} \\ |V_{TP(max)}| + \sqrt{\dfrac{2I_{BP}}{K'_{P(min)}\left(\dfrac{W}{L}\right)_{CB}}} + 0.2\text{ V} \end{cases}$$

$$\approx \text{MAX}\left\{\begin{array}{l} (0.75) + \sqrt{\dfrac{2(2\,\mu)}{(32\,\mu)\left(\dfrac{21}{3.5}\right)}} + \sqrt{\dfrac{2(1\,\mu)}{(80\,\mu)\left(\dfrac{2}{0.35}\right)}} \approx 0.96\text{ V} \\ (0.75) + \sqrt{\dfrac{2(0.5\,\mu)}{(32\,\mu)\left(\dfrac{1}{3}\right)}} + 0.2\text{ V} \approx 1.26\text{ V} \end{array}\right\} \approx 1.26$$

Although combining 3s $K'_{P(min)}$, $K'_{N(min)}$, and $V_{TP(max)}$ values linearly is probabilistically unreasonable, applying $V_{TP(max)}$ alone still violates the $V_{IN(min)}$ specification:

$$V_{IN(min)} \approx |V_{TP(max)}| + \sqrt{\frac{2I_{BP}}{K'_P\left(\dfrac{W}{L}\right)_{CB}}} + 0.2\text{ V} = (0.75) + \sqrt{\frac{2(0.5\,\mu)}{(40\,\mu)\left(\dfrac{1}{3}\right)}} + 0.2\text{ V} \approx 1.22\text{ V}$$

∴ Reduce M_{CB}'s V_{SG} by 150 mV (i.e., decrease $V_{IN(min)}$ to roughly 1.05 V) by (1) decreasing M_{C1}–M_{C2}'s $V_{SD(sat)}$ by 75 mV, (2) decreasing M_{M1}–M_{M2}'s $V_{SD(sat)}$ by 50 mV, and (3) pushing M_{M1}–M_{M2} slightly into triode by 25 mV (since cascode pair M_{C1}–M_{C2} decreases the impact of mirror pair M_{M1}–M_{M2}'s rds's on A_{EA}'s small-signal gain). Also reduce M_{M1}–M_{M2}'s L to $5L_{(min)}$ to keep overall dimensions (i.e., silicon area) from growing excessively.

$$\left(\frac{W}{L}\right)_{C1} \equiv \left(\frac{W}{L}\right)_{C2} \geq \frac{2I_{B2}}{V^2_{SD.C2(sat)}K'_{P(min)}} = \frac{2(1\,\mu)}{(0.075)^2(32\,\mu)} = 11 \quad \text{or} \quad \frac{3.85\,\mu\text{m}}{0.35\,\mu\text{m}}$$

$$\left(\frac{W}{L}\right)_{M1} \equiv \left(\frac{W}{L}\right)_{M2} \geq \frac{2(I_{D1} + I_{B2})}{V^2_{SD.M2(sat)}K'_{P(min)}} = \frac{2(1\,\mu + 1\,\mu)}{(0.1)^2(32\,\mu)} = 12.5 \quad \text{or} \quad \frac{22\,\mu\text{m}}{1.75\,\mu\text{m}}$$

and

$$\left(\frac{W}{L}\right)_{CB} = \frac{2I_{BP}}{K'_P\left[V_{SD.M2} + \sqrt{\frac{2I_{C2}}{K'_P\left(\frac{W}{L}\right)_{C2}}}\right]^2} = \frac{2(0.5\,\mu)}{(40\,\mu)\left[0.075 + \sqrt{\frac{2(1\,\mu)}{(40\mu)\left(\frac{3.85}{0.35}\right)}}\right]^2} = 1.2$$

$$\text{or} \qquad \left(\frac{W}{L}\right)_{CB} \equiv \frac{1.2\,\mu\text{m}}{1\,\mu\text{m}}$$

13. $V_{OS.S}$:

$$V_{OS.S} = \frac{V_{DS.B2} - V_{DS.B1}}{A_{EA}} = \frac{V_{EA} - (V_{IN} - V_{SG.M1})}{A_{EA}}$$

$$= \left\{\begin{matrix} \frac{V_{EA(max)} - (V_{IN} - V_{SG.M1})}{A_{EA}} \\ \frac{V_{EA(min)} - (V_{IN} - V_{SG.M1})}{A_{EA}} \end{matrix}\right\} \equiv \left\{\begin{matrix} V_{OS.S(max)} \\ V_{OS.S(min)} \end{matrix}\right\}$$

where $$V_{OS.S(max)} \le \frac{V_{EA(max)} - \left(V_{IN} - |V_{TP(max)}| - \sqrt{\frac{2I_{M1}}{K'_{P(min)}\left(\frac{W}{L}\right)_{M1}}}\right)}{A_{EA}}$$

$$= \frac{(V_{IN} - 0.3) - \left(V_{IN} - (0.75) - \sqrt{\frac{2(2\,\mu)}{(32\,\mu)\left(\frac{22}{1.75}\right)}}\right)}{(100)} \approx 4.6\text{ mV}$$

and $$V_{OS.S(min)} \le \frac{V_{EA(min)} - \left(V_{IN(max)} - |V_{TP(min)}| - \sqrt{\frac{2I_{M1}}{K'_{P(max)}\left(\frac{W}{L}\right)_{M1}}}\right)}{A_{EA}}$$

$$= \frac{(0.2) - \left((1.6) - (0.75) - \sqrt{\frac{2(2\,\mu)}{(48\,\mu)\left(\frac{22}{1.75}\right)}}\right)}{(100)} \approx -5.7\text{ mV}$$

$\therefore \quad -5.7\text{ mV} = V_{OS.S} = 4.6\text{ mV}.$

14. $V_{OS.S}$'s effect on LDR:

$$\mathrm{LDR}_{OS.S} = \left(\frac{\Delta V_{EA}}{A_{EA}}\right)\left(\frac{V_O}{V_{REF}}\right)$$

$$= \left[\frac{(V_{IN} - V_{SG.PO(min)} + V_{GS.NBUF(min)}) - (V_{IN} - V_{SG.PO(max)} + V_{GS.NBUF(max)})}{A_{EA}}\right]\left(\frac{V_O}{V_{REF}}\right)$$

$$= \left[\frac{(V_{GS.NBUF(min)} - V_{SG.PO(min)}) - (V_{GS.NBUF(max)} - V_{SG.PO(max)})}{A_{EA}}\right]\left(\frac{V_O}{V_{REF}}\right)$$

$$= \left[\sqrt{\frac{2I_{NBUF(min)}}{k'_N\left(\frac{W}{L}\right)_{NBUF}}} - \sqrt{\frac{2I_{O(min)}}{k'_P\left(\frac{W}{L}\right)_{PO}}} - \sqrt{\frac{2I_{NBUF(max)}}{k'_N\left(\frac{W}{L}\right)_{NBUF}}} + \sqrt{\frac{2I_{O(max)}}{k'_P\left(\frac{W}{L}\right)_{PO}}}\right]\left(\frac{1}{A_{EA}}\right)\left(\frac{V_O}{V_{REF}}\right)$$

$$\approx \left(\sqrt{\frac{2(0.5\,\mu)}{(100\,\mu)\left(\frac{12}{0.35}\right)}} - \sqrt{\frac{2(0)}{(40\,\mu)\left(\frac{18.3\,\mathrm{k}}{0.35}\right)}} - \sqrt{\frac{2(46\,\mu)}{(100\,\mu)\left(\frac{12}{0.35}\right)}}\right.$$

$$\left. + \sqrt{\frac{2(50\,\mathrm{m})}{(40\,\mu)\left(\frac{18.3\,\mathrm{k}}{0.35}\right)}}\right)\left(\frac{1}{(100)}\right)\left(\frac{(1)}{(0.9)}\right)$$

$$\approx \left[(17\,\mathrm{m}) - (0) - (164\,\mathrm{m}) + (219\,\mathrm{m})\right]\left[\frac{1}{(100)}\right]\left[\frac{(1)}{(0.9)}\right] \approx 0.8\,\mathrm{mV}$$

15. $V_{OS.R}$:

Assuming transistors laid out to match well (as critical devices) have an equivalent 3σ input-referred mismatch offset v^* of 5 mV:

$$V_{OS.R} \approx \sqrt{v_d^{*2} + v_m^{*2}\left(\frac{g_{m.M}^2}{g_{m.D}^2}\right) + v_b^{*2}\left(\frac{g_{m.B}^2}{g_{m.D}^2}\right)}$$

$$= \sqrt{v_d^{*2} + v_m^{*2}\left[\frac{I_M K'_P\left(\frac{W}{L}\right)_M}{I_D K'_N\left(\frac{W}{L}\right)_D}\right] + v_b^{*2}\left[\frac{I_B\left(\frac{W}{L}\right)_B}{I_D\left(\frac{W}{L}\right)_D}\right]}$$

$$= (5\,\mathrm{m})\sqrt{1 + \left[\frac{(2\,\mu)(40\,\mu)\left(\frac{22}{1.75}\right)}{(1\,\mu)(100\,\mu)\left(\frac{1.75}{3.5}\right)}\right] + \left[\frac{(1\,\mu)\left(\frac{2}{0.35}\right)}{(1\,\mu)\left(\frac{1.75}{3.5}\right)}\right]} \approx 29\,\mathrm{mV}$$

which exceeds the specification by ±4 mV. Had it not been for the increase in $(W/L)_{M1}$ and $(W/L)_{M2}$ to accommodate $V_{IN(min)}$, $V_{OS.R}$ would have been roughly 24 mV. In any case, decreasing $V_{OS.R}$ amounts to increasing $g_{m.D}$ (which increases both A_{EA} and f_{0dB} and compromises stability, in light of A_{BUF}'s finite BW and M_{PO}'s output pole p_O) or decreasing $g_{m.M}$ (which increases $V_{IN(min)}$).

∴ Choose to risk some yield loss (in meeting the $V_{IN\,(min)}$ specification) and/or tighten (i.e., improve) V_{REF}'s accuracy to accommodate a larger $V_{OS.R}$.

General design note: A low $V_{IN(min)}$ design usually trades off gain for stability because compensating for the gain lost as a result of a low-$V_{IN(min)}$ architecture amounts to increasing the number of gain stages in the circuit, which translates to more ac nodes and consequently more poles. Additionally, as the forgoing example shows, a low $V_{IN(min)}$ also increases $V_{OS.R}$. Similarly, low-I_Q operation sacrifices speed because a higher f_{0dB} requires a faster (more power-hungry) buffer. Unfortunately, ensuring the circuit maintains performance across process and temperature corners only exacerbates the effects of these tradeoffs, forcing the designer to balance linear and probabilistic worst-case approaches to assess and accept reasonable risk. In the end, optimal designs, as with most operational amplifiers and analog and mixed-signal ICs, target *reasonable* and *practical* specifications and focus on *key* application-specific performance parameters to avoid the unnecessary sacrifices over-designed circuits dispel while still realizing the value-added features that engage and attract prospective consumers to the market place.

6.4 Summary

Although not always the case, the design cycle of a linear regulator usually starts at the output and ends with the input, with the error amplifier. The load, for instance, determines the type of power transistor needed, the power device defines the parametric limits of the buffer driving it, and the buffer defines those of the preceding input stage. As a whole, taking into account their interdependencies, the entire chain must cater to the increasingly stringent efficiency and accuracy demands of existing and emerging state-of-the-art portable and stationary applications, such as surviving low input supply voltages and requiring low quiescent currents, high loop gains, high bandwidths, low offsets, and other equally important features.

Power PMOS transistors are efficient (and better able to extend battery life) because not only do they accommodate lower input supplies (with lower dropout voltages) but they also demand less current to drive (i.e., gate current is 0 A). N-type devices, on the other hand, are normally more accurate because they respond quicker to transient load-current transitions (i.e., followers yield higher bandwidths), and BJTs, to be more specific, source higher currents and survive higher voltages. Low dropout (LDO) regulators, as a result, frequently use PFETs to conduct and condition power to the load, except applications demanding high accuracy often call for n-type power transistors, just as high-power requirements tend to demand BJTs. Irrespective of the power device used, its physical implementation must ultimately account for the debilitating influence its parasitic components inflict on the circuit. Parasitic vertical and lateral pnp BJTs, for example, and

their constituent diodes not only induce ground current but also compromise the integrity (e.g., proneness to latch-up) and performance (e.g., efficiency) of the IC.

Since power devices are large, they present unique and stringent dc and ac driving requirements. A driving buffer must not only survive low input supplies and use low quiescent currents (as prescribed by the application) but also comply with the wide voltage swings and high-current demands attached to large power transistors. To this end, with respect to the buffer, emitter- and source-follower transistors are compact, fast, and efficient because they generate moderately low output impedances with relatively low currents. Dynamically adjusting their bias points, though, by incorporating localized-feedback loops can decrease the buffer's demand for power during light loading conditions, when the toll of quiescent current on efficiency is greatest. While negative feedback is popular in this regard, positive feedback cannot only boost the gain but also accelerate the response of the buffer, albeit the stability risk it poses to the system. (Note the loop gain of the positive-feedback loop must remain sufficiently below one across the entire spectrum, especially near the regulator's unity-gain frequency, to guarantee stable operating conditions.)

Ultimately, several key specifications converge on the error amplifier, from dc accuracy and forward open-loop gain (which translate to regulation performance) to gain-phase response across frequency and power-supply rejection (PSR). From a design perspective, n-type input differential stages normally favor signal-to-noise ratio (SNR) because their biasing reference voltages are higher, for which injected systematic switching noise constitutes a lower fraction of the reference. While folding loads generally ease headroom requirements (i.e., decrease $V_{IN(min)}$), supply-referenced load mirrors increase (i.e., improve) the PSR of p-type power devices and ground-referenced mirrors that of their n-type counterparts. Since input-referred random offsets depend almost entirely on the transistors in the differential pair and those nearest to the supplies in the first stage (i.e., load mirror and nondegenerated biasing transistors), matching their respective layouts is critical for accuracy performance.

Design is, by nature, the "art of tradeoff" because improving one parameter normally degrades another. Increasing power efficiency by reducing the input supply, for example, not only constrains the circuit to p-type power transistors (for lower dropout) but also increases input-referred offsets and compromises stability. The fact is lower input supplies reduce the margin with which to design the transconductor ratios of critically matched transistors and require folding loads (and extra stages) that necessarily introduce more poles to the feedback system (by way of additional ac nodes). Reducing quiescent currents also costs the system performance in the form of lower speed, higher input-referred offsets (through transconductor ratios), and lower ac accuracy. Worst-case operating conditions and

extreme process and temperature corners further correlate parameters in a way that iteration in the design process is almost impossible to avoid. Managing and assessing the effects of these probabilistic variations on the circuit is an intrinsic part of IC design, without which product development and practical solutions could not viably succeed.

In many ways, the IC design practices and tradeoffs discussed in this chapter embody and fulfill the design objectives of this textbook because they employ the device, circuit, and feedback tools developed in Chaps. 2 through 4 to address and achieve the system and ac design targets discussed in Chaps. 1 and 5. While the text may not cover the entire expanding field of analog microelectronics, the design considerations presented exemplify and describe, to an extent, the core of analog IC design, from device physics, circuit techniques, and positive- and negative-feedback theory to system integration and design approach. The components discussed to this point, however, do not yet constitute a complete linear regulator solution, not without considering (a) the system in light of the circuits that comprise it, (b) possible performance-specific enhancers, (c) protection against extreme operating conditions, and (d) characterization, which the follow-up chapters address.

CHAPTER 7
System Design

The purpose of this chapter is to discuss the assembly and construction of the linear regulator system when targeted for a particular application. To this end, the chapter first combines the transistor-level circuit solutions presented in Chap. 6 to achieve the regulator-specific ac objectives outlined in Chaps. 4 and 5. The chapter then explores how to integrate self-referencing features into the regulator circuit to forego the power and silicon real estate associated with stand-alone voltage references. Some applications may even call for current regulators so the text also reviews and describes circuit strategies for the same. The chapter ends by surveying a series of circuit techniques aimed at optimizing specific performance parameters such as dropout (for low power), accuracy, and power-supply rejection.

7.1 External Compensation

The number and diversity of features consumers enjoy, demand, and in some cases expect from portable, battery-powered devices often outpace technology- and design-driven reductions in energy and power. In other words, with respect to linear regulators targeted for the higher power portable market space, the sum of smaller but considerably many more feature-specific currents remains relatively high at several tens and even hundreds of milliamps. As a result, because many of these features engage and disengage simultaneously at rates faster than their driving supply can respond (because the supply has little to no quiescent current of its own), high output capacitance C_O at regulator output v_O is necessary and external compensation is not only merely justifiable but often also a necessity.

Combining low quiescent-current and low input-voltage requirements, which portable devices demand for high efficiency and extended battery-life operation, with high output capacitance for faster large-signal response typically prompt designers to build externally compensated PMOS regulator solutions, as exemplified by the low-dropout (LDO) circuit shown in Fig. 7.1*a*. The design features an npn differential pair Q_{ND1}-Q_{ND2} loaded with a supply-referenced p-type MOS mirror M_{PM1}-M_{PM2}, an n-type MOS buffer follower M_{NBUF}, and power

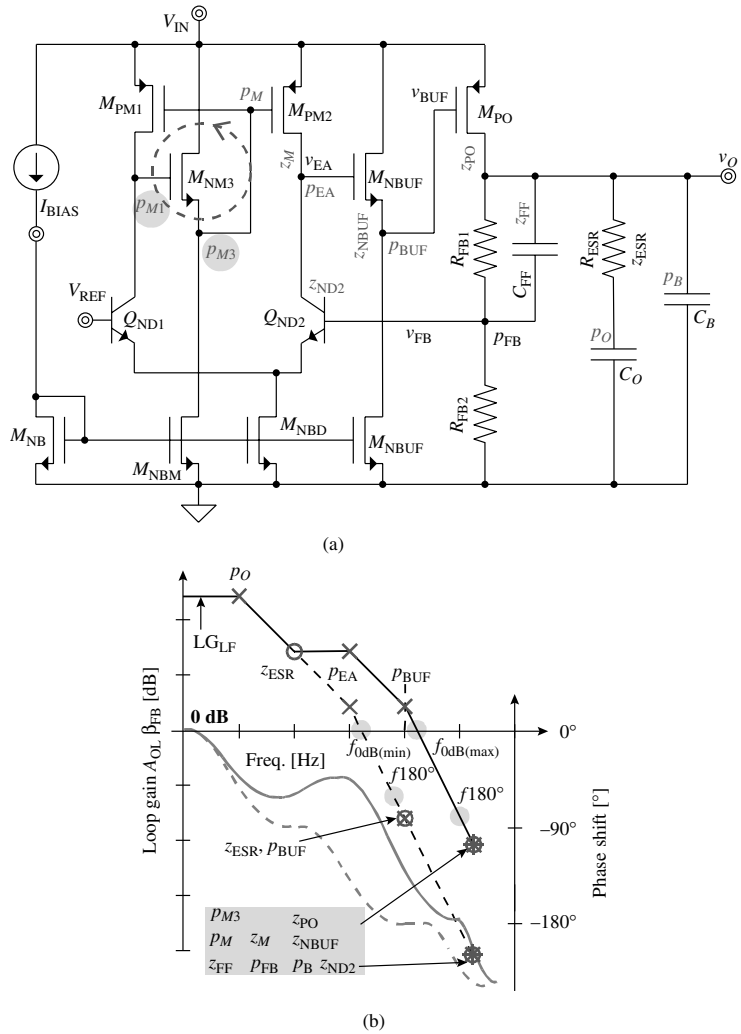


FIGURE 7.1 Externally compensated low-dropout BiCMOS regulator (*a*) circuit and (*b*) corresponding gain-phase response across frequency.

PMOS M_{PO}. N-type follower M_{NM3} reduces dc voltage differences in the mirror (i.e., in $V_{SD.PM1}$ and $V_{SD.PM2}$) and the systematic offsets they produce at the input by matching (to first order) the combined effects of M_{PO}'s source-gate voltage $V_{SG.PO}$ and M_{NBUF}'s gate-source voltage $V_{GS.NBUF}$ to M_{PM1}'s $V_{SG.PM1}$ and M_{NM3}'s $V_{GS.NM3}$. The purpose of capacitor C_{FF} is to feed forward ac signals present at v_O to sense feedback node v_{FB} and, in doing so, introduce a feed-forward left-half-plane zero

(i.e., z_{FF}) to help save phase near the unity-gain frequency of the system (for stability). The input pair transistors are bipolar-junction transistors (BJTs) because they generally match better than metal-oxide-semiconductor (MOS) field-effect transistors (FETs) and consequently produce lower input-referred offsets.

7.1.1 Compensation

The general strategy for compensating an externally compensated LDO regulator is to place output pole p_O at sufficiently low frequencies to dominate over all other poles and zeros in the circuit, as illustrated in Fig. 7.1*b*, and keep unity-gain frequency f_{0dB} well below the parasitic-pole region. In practice, when other poles and zeros remain at higher frequencies across process, temperature, and operating conditions, loop gain drops linearly with frequency past p_O in a single-pole roll-off manner, roughly setting f_{0dB}'s minimum point when low-frequency loop gain LG_{LF} reaches 0 dB after traversing a gain-bandwidth product GBW (i.e., $LG_{LF}\, p_O$) expanse of frequency:

$$\frac{LG_{LF}}{1+\frac{2\pi s}{p_O}} \approx \left.\frac{LG_{LF}}{\left(\frac{2\pi s}{p_O}\right)}\right|_{f_{0dB(min)} \approx LG_{LF} p_O \equiv GBW} \equiv 1 \tag{7.1}$$

Note $f_{0dB(min)}$ is often a specification target.

Building gain in error amplifier A_{EA} requires a high-resistance node so the pole associated with that node (i.e., p_{EA}) is normally the second dominant pole in the circuit. Although the equivalent series resistance (ESR) of output capacitor C_O introduces what could be a phase-saving zero z_{ESR} to the response near p_{EA}, R_{ESR}, and consequently z_{ESR}'s location vary substantially across process and temperature:

$$\frac{1}{2\pi R_{ESR(max)} C_O} \le z_{ESR} \le \frac{1}{2\pi R_{ESR(min)} C_O} \tag{7.2}$$

As a result, placing p_{EA} well below $f_{0dB(min)}$ and expecting z_{ESR} to cancel p_{EA}'s effects is risky, but restraining $f_{0dB(max)}$ below $z_{ESR(min)}$ in all cases, on the other hand, is unfavorably pessimistic. The aim is to therefore place second dominant pole p_{EA} near $f_{0dB(min)}$, well below $f_{0dB(max)}$:

$$\begin{aligned} p_{EA} &\approx \frac{1}{2\pi(r_{ds.MP2} \parallel r_{o.ND2})(C_{DG.MP2} + C_{DB.MP2} + C_{\mu.ND2} + C_{DB.MNBUF})} \\ &\approx f_{0dB(min)} \end{aligned} \tag{7.3}$$

To keep p_{EA} near $f_{0dB(min)}$, the resistance at v_{EA} must be moderately low so the channel lengths of mirror devices M_{PM1} and M_{PM2} are relatively short (to allow channel-length modulation λ to decrease M_{PM2}'s small-signal output resistance $r_{ds.PM2}$). Note $z_{ESR(min)}$ extends $f_{0dB(min)}$ by how far $z_{ESR(min)}$ precedes p_{EA}:

$$f_{0dB(max)} \approx f_{0dB(min)}\left(\frac{p_{EA}}{z_{ESR(min)}}\right) \tag{7.4}$$

As frequencies increase, the next pole to appear is buffer pole p_{BUF}, and since p_O and p_{EA} already shift considerable phase in the absence of z_{ESR}, p_{BUF} must remain well above $f_{0dB(min)}$, but not necessarily higher than $f_{0dB(max)}$. In fact, since z_{ESR} is the reason f_{0dB} shifts to $f_{0dB(max)}$ in the first place, z_{ESR} offsets the effects of p_O and adds margin for p_{BUF} to reside near $f_{0dB(max)}$:

$$p_{BUF} \approx \frac{g_{m.NBUF}}{2\pi C_{SG.PO}} \approx f_{0dB(max)} \tag{7.5}$$

After p_{BUF}, there is no margin left for the remaining parasitic poles in the system, which include feedback sense pole p_{FB}, bypass pole p_B, and mirror pole p_M, so they must remain well above $f_{0dB(max)}$.

Unfortunately, keeping parasitic poles at high frequencies demands power in the form of additional quiescent current, which is why shifting them just high enough beyond $f_{0dB(max)}$ for stability conditions to remain true is the general design approach. Although the effect of each pole a decade past $f_{0dB(max)}$ on phase margin is minimal, their relatively small but compounded contributions may still have an avalanching effect on phase response near f_{0dB}. Regrettably, a rapidly decreasing phase near f_{0dB} implies the circuit's proneness to unstable conditions is more sensitive to process and temperature variations because relatively small shifts in pole and zero locations amount to larger changes in phase, which means the circuit is conditionally stable and may oscillate during start-up and/or in response to quick load dumps.

Fortunately, nearby left-half-plane zeros dampen the avalanching effects the remaining poles have on phase response by recovering some of the phase lost. The general design tactic is to keep parasitic poles barely a decade above $f_{0dB(max)}$ and introduce, when possible, just as many left-half-plane zeros to offset them. Feed-forward capacitor C_{FF} presents such a phase saving zero (at z_{FF}) by shorting feedback resistor R_{FB1} past $1/2\pi R_{FB1}C_{FF}$:

$$\left.\frac{1}{sC_{FF}}\right|_{z_{FF}=\frac{1}{2\pi R_{FB1}C_{FF}}\approx 5 f_{0dB(max)}} \equiv R_{FB1} \tag{7.6}$$

Introducing a zero below $f_{0dB(max)}$ in the presence of z_{ESR} however, extends $f_{0dB(max)}$ closer to the parasitic-pole region, where the now

closer parasitic poles have a more pronounced impact on phase margin. So, to prevent process and temperature variations from pulling z_{FF} below $f_{0dB(max)}$, z_{FF} is placed, by design, slightly above $f_{0dB(max)}$.

Now that z_{FF} adds some margin to the system, parasitic bypass and feedback poles p_B and p_{FB} can now comfortably reside barely a decade above $f_{0dB(max)}$. Recall from Chap. 5 that the shunting effect of C_B on R_{ESR} produces p_B:

$$p_B \approx \frac{1}{2\pi R_{ESR} C_{B(max)}} \approx 10 f_{0dB(max)} \tag{7.7}$$

so keeping p_B at $10f_{0dB(max)}$ bounds bypass capacitance below $C_{B(max)}$, which in many cases is not a problem because C_B only comprises parasitic on-chip and printed-circuit-board (PCB) trace capacitance. Similarly, the capacitance present at feedback sense signal v_{FB} confines R_{FB1} and R_{FB2}'s combined parallel resistance because the capacitance should not short $R_{FB1} \,||\, R_{FB2}$ below $10f_{0dB(max)}$:

$$\left.\frac{1}{s(C_{\mu.ND2} + C_{\pi.ND2} + C_{FF})}\right|_{p_{FB}=\frac{1}{2\pi(R_{FB1} \,||\, R_{FB2})(C_{\mu.ND2}+C_{\pi.ND2}+C_{FF})}\approx 10 f_{0dB(max)}}$$

$$\equiv R_{FB1} \,||\, R_{FB2} \tag{7.8}$$

where Q_{ND2}'s multiplying Miller effect on $C_{\mu.ND2}$ is negligible because the parasitic capacitance present at Q_{ND2}'s collector (i.e., v_{EA}) already shorted the resistance at v_{EA} and the gain across Q_{ND2} is consequently low at these frequencies. Unlike p_B, though, which often stays at high frequencies without much effort, p_{FB} tends to creep into lower frequencies because feedback resistances R_{FB1} and R_{FB2} are usually high to keep the quiescent current flowing through them relatively low at less than 1–2 μA.

Load mirror M_{PM1}-M_{PM2} generally introduces a closely spaced pole-zero pair so its collective impact on phase is not as pronounced. Level-shifting follower M_{NM3} and mirroring transistor M_{PM1} establish a negative-feedback loop that must be, by itself, stable. This inner loop has two poles: p_{M1} and p_{M3} at M_{NM3}'s gate and source terminals, respectively. Pole p_{M1} must be sufficiently high to ensure the open-loop unity-gain frequency of the inner loop (i.e., $f_{0dB.M}$), which is the loop's gain-bandwidth product (i.e., $A_{V.LF}p_{M1}$) and the equivalent closed-loop bandwidth of the mirror (i.e., p_M), is slightly above $f_{0dB(max)}$:

$$p_M \equiv f_{0dB.M} \approx A_{V.LF} p_{M1} \approx \frac{g_{m.PM1}(r_{ds.PM1} \,||\, r_{o.ND1})}{2\pi(r_{ds.PM1} \,||\, r_{o.ND1})C_{G.NM3}}$$

$$= \frac{g_{m.PM1}}{2\pi C_{G.NM3}} \approx \frac{g_{m.PM1}}{2\pi(C_{GD.NM3} + C_{DB.PM1} + C_{DG.PM1} + C_{\mu.ND1})} \approx 5 f_{0dB(max)} \tag{7.9}$$

where p_M is the negative-feedback translation of p_{M1} (i.e., negative feedback shifts bandwidth-limiting p_{M1} to $f_{0dB.M}$) and $C_{G.NM3}$ is the total parasitic capacitance present at the gate of M_{NM3}.

To ensure the mirror loop is stable, M_{NM3}'s gate pole p_{M1} must be dominant and p_{M3} must reside near closed-loop bandwidth $f_{0dB.M}$ (i.e., close to p_M):

$$p_{M3} \approx \frac{g_{m.NM3}}{2\pi(C_{SG.PM1} + C_{SG.PM2} + C_{DG.PM2} + C_{SG.NM3})} \approx f_{0dB.M}$$
$$= p_M \approx 5f_{0dB(max)} \tag{7.10}$$

where M_{P2}'s Miller-multiplying effect on $C_{DG.PM2}$ is negligible because, like before, p_{EA} already shorted the resistance at v_{EA} to ac ground, forcibly reducing the common-source gain across M_{P2} at these higher frequencies. Collectively, the loop introduces the feedback translation of p_{M1} as mirror pole p_M to the system and, when ac signals from input device Q_{ND2} overwhelm those sourced by load mirror M_{PM2}, the mirror introduces mirror zero z_M at roughly $2p_M$, which as before, helps reduce the avalanching effects this and other poles in the vicinity have on phase response.

In-phase feed-forward capacitor $C_{GS.NBUF}$ in n-type voltage following buffer M_{NBUF} also introduces a left-half-plane zero (i.e., z_{NBUF}), except its location is normally high because M_{NBUF} and consequently $C_{GS.NBUF}$ are relatively small:

$$i_{C.NBUF} = \left.\frac{v_{gs.NBUF}}{\left(\frac{1}{sC_{GS.NBUF}}\right)}\right|_{z_{NBUF} \approx \frac{g_{m.NBUF}}{2\pi C_{GS.NBUF}} \geq 5f_{0dB(max)}} \equiv i_{gm.NBUF} = v_{gs.NBUF}g_{m.NBUF} \tag{7.11}$$

Similarly, out-of-phase feed-forward capacitor $C_{GD.PO}$ in power PMOS M_{PO} also introduces a zero (i.e., z_{PO}), but this time on the right half of the s plane, which means z_{PO} adds gain but subtracts phase (like a pole) so it must remain at or above $10f_{0dB(max)}$. Fortunately, z_{PO} tends to naturally reside at high frequencies because M_{PO}'s large transconductance (given M_{PO} is a large power device carrying high current) offsets the polarity-reversing current contribution of $C_{GD.PO}$:

$$i_{CGD} = \left.\frac{v_{gs.PO} - v_{ds.PO}}{\left(\frac{1}{sC_{GD.PO}}\right)}\right|_{z_{PO}} = \left.\frac{v_{gs.PO}}{\left(\frac{1}{sC_{GD.PO}}\right)}\right|_{z_{PO} \approx \frac{g_{m.PO}}{2\pi C_{GD.PO}} \geq 10f_{0dB(max)}}$$
$$\equiv i_{gm.PO} = v_{gs.PO}g_{m.PO} \tag{7.12}$$

Differential input-pair transistor Q_{ND2}'s Miller capacitor $C_{\mu.ND2}$ also feeds forward out-of-phase ac signals across Q_{ND2} and therefore introduces right-half-plane zero z_{ND2}. Like z_{PO}, z_{ND2} must remain at or above $10f_{0dB(max)}$, but like z_{NBUF}, z_{ND2} naturally gravitates to high frequencies because Q_{ND2} and its $C_{\mu.ND2}$ are relatively small (and Q_{ND2}'s transconductance $g_{m.ND2}$ is relatively high because Q_{ND2}'s collector current varies exponentially with respect to base-emitter voltage):

$$i_{C\mu} = \left.\frac{v_{be.ND2} - v_{ce.ND2}}{\left(\frac{1}{sC_{\mu.ND2}}\right)}\right|_{z_{ND2}} = \left.\frac{v_{be.ND2}}{\left(\frac{1}{sC_{\mu.ND2}}\right)}\right|_{z_{ND2} \approx \frac{g_{m.ND2}}{2\pi C_{\mu.ND2}} \geq 10 f_{0dB(max)}}$$

$$\equiv i_{gm.ND2} = v_{be.ND2} g_{m.ND2} \tag{7.13}$$

7.2 Internal Compensation

The need for higher power levels and quicker response times are as imminent as the demand for higher integration, which is where internal compensation schemes thrive. The fact is packing more features into a single IC introduces uncorrelated noise into the supply, reducing, as a result, the signal-to-noise-ratio (SNR) performance of sensitive analog electronics attached to that supply. Dedicated on-chip regulators help in this respect because they decouple the otherwise common noise from sensitive loads. Having no external pad or pin to rely on, though, the challenge in designing these point-of-load (PoL) regulators is the unavailability of off-chip capacitors. Fortunately, the lower power levels these targeted loads demand (and their smaller large-signal changes in load) offset some of the expense (i.e., silicon real estate) associated with on-chip capacitors, as lower capacitances may now satisfy the relatively modest needs of a lighter load.

Low-dropout performance in dedicated on-chip supplies, as it turns out, is not always a requirement because the power dissipated by the rest of the system, on average, often overwhelms that of the particular load in question. If other considerations allow, using this argument to ease dropout requirements is important because higher dropout n-type power devices outperform their lower dropout p-type counterparts in speed and consequently in on-chip capacitance silicon-area requirements. Headroom, by the way, may sometimes constrain dropout voltages to low values, like when one switching regulator supplies the entire IC and its output is already low to mitigate the losses associated with the high-power sectors of the system.

7.2.1 High-Dropout Regulator

Considering low quiescent current is always an overriding concern, the high-dropout (HDO) regulator illustrated in Fig. 7.2*a* employs an NMOS power device (i.e., M_{NO}) because the gate draws no dc current. The design also features an n-type differential MOS pair M_{ND1}-M_{ND2} folded with cascode pair M_{PC1}-M_{PC2} into ground-referenced mirror M_{NM1}-M_{NM2}. Compensation capacitor C_C helps pull dominant error-amplifier pole p_{EA} to lower frequencies by shunting small ac signals to ground and C_C and M_{NB1}'s diode-connected $1/g_{m.NB1}$ resistor introduce left-half-plane zero z_C to offset the impact of output pole p_O. Using M_{NB1} to introduce a zero is more compelling than placing an additional in-phase feed-forward capacitor C_{FF} across R_{FB1} because (1) C_{FF} requires additional silicon area and (2) $1/g_{m.NB1}$ tracks the $1/g_{m.NO}$ resistance in output pole p_O better with respect to process and temperature. Capacitor C_C is directed to

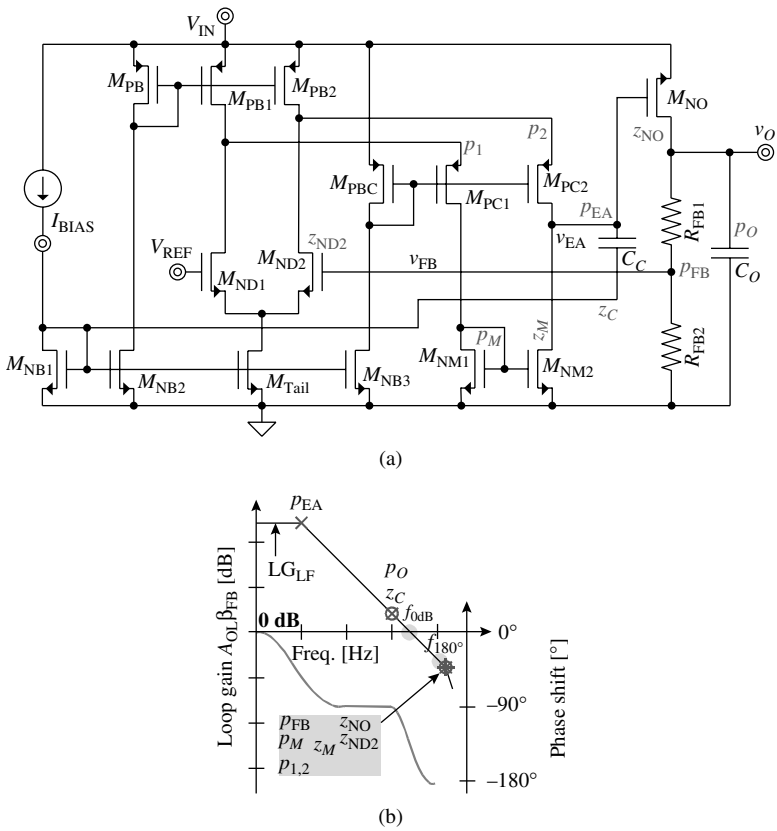


Figure 7.2 Internally compensated high-dropout (HDO) CMOS regulator (*a*) circuit and (*b*) corresponding gain-phase response across frequency.

ground (instead of input supply v_{IN}) to avoid coupling input-supply ripple v_{in} to v_O via M_{NO}, which is in a follower configuration. Output capacitor C_O (even if only at the picofarad range) helps source and sink high-frequency load current past the reach of M_{NO}'s bandwidth (and the bandwidth of the outer loop that supports it).

Compensation

The general compensation strategy in an internally compensated regulator is to (1) set a dominant low-frequency pole, (2) place the second dominant pole within a decade of unity-gain frequency f_{0dB}, (3) add a left-half-plane zero at or slightly above the second pole to offset the effects of the second pole, and (4) push all parasitic poles within a decade above f_{0dB}, keeping natural left-half-plane zeros in the vicinity to prevent the phase response from avalanching. Error-amplifier pole p_{EA} is typically dominant over all other poles and zeros, as shown in Fig. 7.2*b*, because the resistance of the gain-setting node is inherently high and all other nodes, including the output, have relatively lower resistances at roughly $1/g_m$. Like in the LDO case, but now with p_{EA}, loop gain drops from its low-frequency point of LG_{LF} to 0 dB at gain-bandwidth product $LG_{LF}p_{EA}$ or $g_{m.ND}/2\pi(C_C + C_{DG.NO})$:

$$\frac{LG_{LF}}{1+\frac{2\pi s}{p_{EA}}} \approx \frac{LG_{LF}}{\left(\frac{2\pi s}{p_{EA}}\right)} \approx \frac{g_{m.ND}r_{ds.NM2}}{r_{ds.NM2}(C_C + C_{DG.NO})s}\bigg|_{f_{0dB}\approx\frac{g_{m.ND}}{2\pi(C_C+C_{DG.NO})}} \equiv 1 \qquad (7.14)$$

where f_{0dB} is often a specification target.

Output capacitance C_O is intentionally as high as silicon or PCB real estate allows because C_O momentarily supplies all the load current power transistor M_{NO} is not fast enough to provide, which means the pole C_O produces tends to dominate over all remaining parasitic poles in the circuit. As a result, C_O together with M_{NO}'s $C_{GS.NO}$ produce the effects of output pole p_O when they shunt M_{NO}'s output resistance $1/g_{m.NO}$:

$$\frac{1}{s(C_O + C_{GS.NO})}\bigg|_{p_{O(min)}\approx\frac{g_{m.NO}}{2\pi(C_{O(max)}+C_{GS.NO})}\geq\frac{f_{0dB}}{10}} \equiv \frac{1}{g_{m.NO}} \qquad (7.15)$$

where $C_{GS.NO}$ helps shunt $1/g_{m.NO}$ to ground through the already shorted capacitive impedance at v_{EA} (because p_{EA} shorts v_{ea} to ac ground at lower frequencies). The left-half-plane zero C_C-M_{NB1} introduce (i.e., z_C), which appears when C_C's impedance is equal to or smaller than M_{NB1}'s $1/g_{m.NB1}$ resistance and the total impedance through the series C_C-M_{NB1} network reduces to $1/g_{m.NB1}$, is meant to cancel the effects of p_O, allowing p_O (and requiring z_C) to precede f_{0dB} slightly:

$$\frac{1}{sC_C}\bigg|_{z_C=\frac{g_{m.NB1}}{2\pi C_C}\equiv p_O} \equiv \frac{1}{g_{m.NB1}} \qquad (7.16)$$

Matching and correlating p_O and z_C across process, and worse of all, load is difficult because p_O, through M_{NO}'s $g_{m.NO}$, depends on output current I_O. A two-decade variation in I_O, such as 10 µA to 1 mA, just to cite an example, produces a one-decade shift in $g_{m.NO}$ and p_O (because $g_{m.NO}$ increases with the square root of I_O), considering mismatches in C_C and $C_{GS.NO} + C_O$ exacerbate differences in z_C and p_O. As a result, in case p_O precedes z_C by almost one decade, stability concerns constrain p_O (and C_O) to remain roughly within a decade of f_{0dB} (i.e., $p_{O(min)}$ must be greater than $0.1f_{0dB}$). Although not immediately obvious, fully dipping the phase to 180° below f_{0dB} (before z_C recovers it), even if only briefly, increases the likelihood f_{0dB} crosses the 180° phase-shift frequency during start-up conditions, while poles, zeros, and f_{0dB} transition to their final locations.

Past p_{EA}, p_O, and z_C, there is slightly less than 90° of phase margin left so, to keep f_{0dB} as high as possible without losing considerable quiescent current, parasitic poles may reside only within one decade above f_{0dB} if left-half-plane zeros hover nearby to prevent the phase response from avalanching. One of the first parasitic poles to appear after p_O is the one at feedback signal v_{FB} because the resistance there ($R_{FB1} \,||\, R_{FB2}$) is often moderately high (and low resistances would otherwise induce higher quiescent current through R_{FB1} and R_{FB2}) and the capacitance present is moderate with at least one C_{GS}. As a result, feedback pole p_{FB} asserts its influence when $C_{GD.ND2}$ and $C_{GS.ND2}$ shunt $R_{FB1} \,||\, R_{FB2}$:

$$\left.\frac{1}{s(C_{GD.ND2} + C_{GS.ND2})}\right|_{p_{FB} = \frac{1}{2\pi(R_{FB1} \,||\, R_{FB2})(C_{GD.ND2} + C_{GS.ND2})} \approx 10 f_{0dB(max)}} \equiv R_{FB1} \,||\, R_{FB2} \tag{7.17}$$

where M_{ND2}'s Miller effect on $C_{GD.ND2}$ is negligible because there is little gain across M_{ND2} and R_{FB1} is in parallel with R_{FB2} because p_O, at these frequencies, already shorted output v_o to ac ground.

All other resistances in the circuit are near $1/g_m$ so the next poles to consider are mirror pole p_M and folding poles p_1 and p_2 because their respective capacitances include at least one C_{GS} (and C_{GS} and C_π are usually considerably larger than C_{GD} and C_μ). Mirror pole p_M results when $C_{GS.NM1}$, $C_{GS.NM2}$, and $C_{DG.NM2}$ shunt the $1/g_{m.NM1}$ resistance M_{NM1} presents at approximately $g_{m.NM1}/2\pi(C_{GS.NM1} + C_{GS.NM2})$:

$$\left.\frac{1}{s(C_{GS.NM1} + C_{GS.NM2} + C_{GD.NM2})}\right|_{p_M \approx \frac{g_{m.NM1}}{2\pi(C_{GS.NM1} + C_{GS.NM2} + C_{GD.NM2})} \approx 5 f_{0dB}} \equiv \frac{1}{g_{m.NM1}} \tag{7.18}$$

except mirror zero z_M offsets p_M's effects soon after at $2p_M$, which is why its placement can remain only slightly above f_{0dB}. Folding poles

p_1 and p_2 at the source terminals of cascode transistors M_{PC1} and M_{PC2} occur when $C_{GS.PC}$, $C_{DG.PB}$, and $C_{GD.ND}$ short the equivalent $1/g_{m.PC}$ resistances present:

$$\left.\frac{1}{s(C_{GS.PC}+C_{DG.PB}+C_{GD.ND})}\right|_{p_{1,2}\approx\frac{g_{m.PC}}{2\pi(C_{GS.PC}+C_{DG.PB}+C_{GD.ND})}\approx 10f_{0dB}} \equiv \frac{1}{g_{m.PC}} \quad (7.19)$$

where the resistance into M_{PC2}'s source is unaffected by the $r_{ds.NM2}$ load because, at these frequencies, p_{EA} already shorted $r_{ds.NM2}$ to ac ground. Although these poles only include one C_{GS}, unlike p_M, which has two, they have no zero to help offset their effects so they should reside a decade above f_{0dB}. Note p_1 and p_2 affect each half of the differential signal equally so their collective effect on the full differential response is one pole at $p_{1,2}$.

Although power device M_{NO}'s in-phase feed-forward capacitor $C_{GS.NO}$ is relatively large, the left-half-plane zero (i.e., z_{NO}) $C_{GS.NO}$ introduces tends to reside at high frequencies because $C_{GS.NO}$'s current must overwhelm the relatively large transconductor current M_{NO} sources before having any impact on the response:

$$i_{C.GS} = \left.\frac{v_{gs.NO}}{\left(\frac{1}{sC_{GS.NO}}\right)}\right|_{z_{NO}=\frac{g_{m.NO}}{2\pi C_{GS.NO}}\geq 5f_{0dB}} \equiv i_{gm.NO} = v_{gs.NO}g_{m.NO} \quad (7.20)$$

Nevertheless, ensuring z_{NO} remains slightly above f_{0dB} to avoid unnecessarily extending f_{0dB} to the parasitic-pole region is important, unless C_O is negligibly small, in which case z_{NO} cancels p_O and z_C should stay slightly above f_{0dB}. The zero M_{ND2}'s out-of-phase feed-forward capacitor $C_{GD.ND2}$ introduces when $C_{GD.ND2}$'s current exceeds $g_{m.ND2}$'s current, however, must in all cases remain at least one decade above f_{0dB} because z_{ND2} resides on the right half of the s plane:

$$i_{C.GD} = \left.\frac{v_{gs.ND2}-v_{ds.ND2}}{\left(\frac{1}{sC_{GD.ND2}}\right)}\right|_{z_{ND2}} = \left.\frac{v_{gs.ND2}}{\left(\frac{1}{sC_{GD.ND2}}\right)}\right|_{z_{ND2}\approx\frac{g_{m.ND2}}{2\pi C_{GD.ND2}}\geq 10f_{0dB(max)}}$$

$$\equiv i_{gm.ND2} = v_{gs.ND2}g_{m.ND2} \quad (7.21)$$

Fortunately, $C_{GD.ND2}$ is typically not large so placing z_{ND2} at higher frequencies is not terribly difficult.

7.2.2 Low-Dropout Regulator

As in the high-power low-dropout (LDO) case described earlier, an LDO topology uses a MOS power transistor to forego the base-current requirements of a BJT and the device is p-type, as shown in the sample embodiment of Fig. 7.3*a*, to reduce dropout voltage. The particular design shown employs n-type differential pair M_{ND1}-M_{ND2} loaded with folding cascoded mirror M_{PM1}-M_{PM2} and corresponding biasing devices M_{NB1} and M_{NB2}. Compensation Miller capacitor C_C pulls error amplifier pole p_{EA} to lower frequencies and nulling resistor R_N pulls power transistor M_{PO}'s right-half-plane zero to the left-half-plane zero (as z_N). As before, C_O helps supply and sink high-frequency load current.

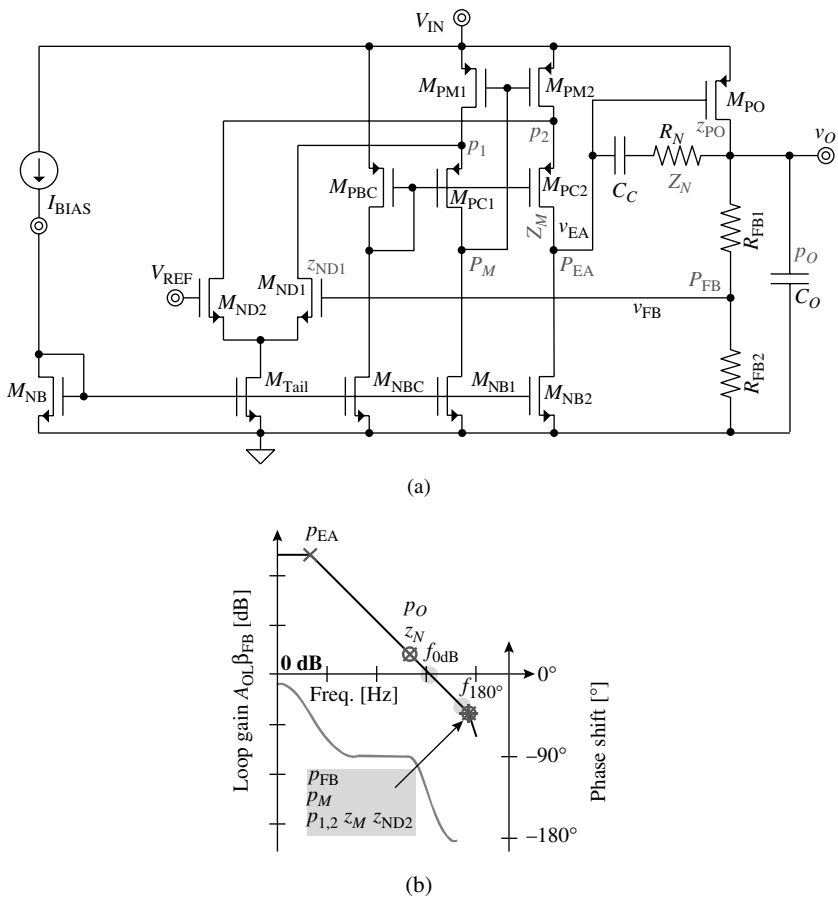


Figure 7.3 Internally compensated low-dropout CMOS regulator (*a*) circuit and (*b*) corresponding gain-phase response across frequency.

Compensation

Although the LDO topology shown in Fig. 7.3 seems vastly different from the HDO counterpart in Fig. 7.2, the compensation strategy is surprisingly similar. As in the HDO case, error amplifier pole p_{EA} is dominant, except now Miller capacitor C_C exploits the inverting gain across power PMOS device M_{PO} (i.e., A_{PO}) to shunt biasing transistor M_{NB2}'s $r_{ds.NB2}$ at p_{EA}. Low-frequency gain LG_{LF} therefore drops linearly past p_{EA} until it reaches 0 dB, in the absence of other poles and zeros below f_{0dB}, at gain-bandwidth product $LG_{LF}p_{EA}$ or $g_{m.ND}/2\pi(C_C + C_{DG.PO})$:

$$\frac{LG_{LF}}{1+\frac{2\pi s}{p_{EA}}} \approx \frac{LG_{LF}}{\left(\frac{2\pi s}{p_{EA}}\right)}$$

$$\approx \frac{g_{m.ND}r_{ds.NM2}A_{PO}}{r_{ds.NM2}[A_{PO}(C_C + C_{DG.PO})]s}\bigg|_{f_{0dB}\approx\frac{g_{m.ND}}{2\pi(C_C+C_{DG.PO})}} \equiv 1 \qquad (7.22)$$

Output pole p_O then follows because output capacitance C_O is considerably high (as much as the application allows) to accommodate high-frequency load dumps. This time the effects of p_O assert their influence on the loop when C_O and M_{PO}'s $C_{DB.PO}$ shunt the $1/g_{m.PO}$ diode-connected resistance that results after C_C and $C_{DG.PO}$ short M_{PO}'s gate-drain terminals, which happens at lower frequencies, past p_{EA}:

$$\frac{1}{s(C_O + C_{DB.PO})}\bigg|_{p_O\approx\frac{g_{m.PO}}{2\pi(C_O+C_{DB.PO})}\geq\frac{f_{0dB}}{10}} \equiv \frac{1}{g_{m.PO}} \qquad (7.23)$$

Nulling resistance is high enough to not only block out-of-phase feed-forward currents through C_C but also shift the feed-forward zero to the left half of the s plane and help offset the effects of p_O. Recall z_N occurs when nulling path current i_n equals and surpasses M_{PO}'s i_{gm}:

$$i_n = \frac{(v_{gs.PO} - v_{ds.PO})}{R_N + \frac{1}{sC_C}}\bigg|_{z_N} = \frac{v_{gs.PO}}{R_N + \frac{1}{sC_C}}\bigg|_{z_N\approx\frac{1}{2\pi\left[R_N-\left(\frac{1}{g_{m.PO}}\right)\right]C_C}\approx p_O}$$

$$\equiv i_{gm} = v_{gs.PO}g_{m.PO} \qquad (7.24)$$

Unfortunately, as in the HDO counterpart, p_O shifts considerably with output current I_O while z_N does not; and to make matters worse, z_N does

not even track p_O with process or temperature. As a result, p_O (and z_N) should not precede f_{0dB} by more than a decade to avoid the phase shift across the loop from dipping to 180° at any frequency below f_{0dB}.

The parasitic poles and zeros that appear past p_{EA}, p_O, and z_N, as before, reside only within one decade past f_{0dB} and left-half-plane zeros hover nearby to prevent the phase response from avalanching. The first pole to consider past p_O and z_N is, again, feedback sense pole p_{FB} because $R_{FB1} \| R_{FB2}$ is moderately high, by design, to reduce quiescent current flow through R_{FB1} and R_{FB2}:

$$\left.\frac{1}{s(C_{GD.ND1}+C_{GS.ND1})}\right|_{p_{FB}=\frac{1}{2\pi(R_{FB1}\|R_{FB2})(C_{GD.ND1}+C_{GS.ND1})}\approx 10f_{0dB(max)}} \equiv R_{FB1}\|R_{FB2} \tag{7.25}$$

where, as before, the Miller effects of $C_{GD.ND1}$ are negligibly small because there is little gain across M_{ND1} and R_{FB1} is in parallel with R_{FB2} because p_O already shorted v_o to ac ground at these frequencies.

As in the HDO case, mirror and folding poles p_M and $p_{1,2}$ are next because the remaining resistances in the circuit are roughly $1/g_m$ and p_M and $p_{1,2}$ include C_{GS} capacitances, which are moderately high with respect to C_{GD} capacitances. Mirror pole p_M occurs when $C_{SG.PM1}$, $C_{SG.PM2}$, $C_{DG.PM1}$, and $C_{DG.PM2}$ short M_{PM1}'s $1/g_{m.PM1}$ resistance:

$$\left.\frac{1}{s(C_{SG.PM1}+C_{SG.PM2}+C_{GD.PM1}+C_{GD.PM2})}\right|_{p_M\approx\frac{g_{m.PM1}}{2\pi(C_{SG.PM1}+C_{SG.PM2}+C_{GD.PM1}+C_{GD.PM2})}\approx 5f_{0dB}} \equiv \frac{1}{g_{m.PM1}} \tag{7.26}$$

except mirror zero z_M offsets its effects soon after at $2p_M$, which is why p_M's placement may remain slightly above f_{0dB}. Similar to the externally compensated LDO, cascode current buffer M_{PC1} closes the mirror's local feedback loop and the pole at M_{PC1}'s source should therefore coincide or exceed p_M to avoid exposing this loop to unstable conditions. This condition is automatically met when considering folding pole p_2 must remain a decade above f_{0dB} because no left-half-plane zeros exist to offset its effects. As in the HDO circuit, the influence of $p_{1,2}$ is felt when $C_{SG.PC}$, $C_{DG.ND}$, and $C_{DG.PM}$ shunt the $1/g_{m.PC}$ resistance present at M_{PC1}-M_{PC2}'s respective sources:

$$\left.\frac{1}{s(C_{SG.PC}+C_{DG.PM}+C_{DG.ND})}\right|_{p_{1,2}\approx\frac{g_{m.PC}}{2\pi(C_{SG.PC}+C_{DG.PM}+C_{DG.ND})}\approx 10f_{0dB}} \equiv \frac{1}{g_{m.PC}} \tag{7.27}$$

where the loading influence of $r_{ds.NB2}$ on M_{PC2}'s source $1/g_{m.PC2}$ resistance is negligible because the Miller capacitance present at v_{EA} shorts $r_{ds.NB2}$ to ac ground at lower frequencies.

In the HDO regulator previously shown, in-phase feed-forward capacitor $C_{GS.NO}$ introduces an additional left-half-plane zero to the mix that allows f_{0dB} to drift slightly to higher frequencies. The internally compensated LDO circuit illustrated in Fig. 7.3 unfortunately does not enjoy this luxury because its nulling zero z_N is already dedicated to canceling p_O's effects. Like z_{FF} in the externally compensated LDO case, however, a feed-forward capacitor across R_{FB1} can introduce another left-half-plane zero, but doing so sacrifices silicon real estate (i.e., money). In any case, the last zero to consider is the one out-of-phase feed-forward capacitor $C_{GD.ND1}$ introduces when its current overwhelms M_{ND1}'s transconductor current $i_{gm.ND1}$:

$$i_{C.GD} = \left.\frac{v_{gs.ND1} - v_{ds.ND1}}{\left(\frac{1}{sC_{GD.ND1}}\right)}\right|_{z_{ND1}} = \left.\frac{v_{gs.ND1}}{\left(\frac{1}{sC_{GD.ND1}}\right)}\right|_{z_{ND1} \approx \frac{g_{m.ND1}}{2\pi C_{GD.ND1}} \geq 10 f_{0dB(max)}}$$
$$\equiv i_{gm.ND1} = v_{gs.ND1} g_{m.ND1} \tag{7.28}$$

and its location should exceed f_{0dB} by at least one decade, which is normally not difficult to accomplish because $C_{GD.ND1}$ is relatively small. Note M_{PO}'s intrinsic out-of-phase feed-forward parasitic capacitor $C_{GD.MPO}$ does not and cannot include a current-limiting resistor like R_N so $C_{GD.MPO}$ introduces another right-half-plane zero z_{PO} to the loop. Fortunately, z_{PO} is at considerably higher frequencies because the transconductor current $C_{GD.MPO}$'s displacement current must overwhelm to invert the phase across M_{PO} is considerably high (i.e., M_{PO}'s $i_{gm.PO}$ is relatively high).

What is interesting and perhaps useful in a Miller-compensated LDO (as shown) is the circuit's tendency to remain stable across a wide range of values of output capacitance C_O. The case thus far presented assumes C_O is relatively small and bounded to keep p_O within a decade of f_{0dB}, but consider that increasing C_O pulls p_O to lower frequencies while simultaneously pushing p_{EA} to higher frequencies. Pole p_{EA} shifts because shunting v_o to ac ground reduces the gain across M_{PO} and the Miller impact M_{PO} has on C_C and $C_{DG.PO}$. As a result, if C_O is sufficiently large, p_O becomes dominant because C_O and $C_{DB.PO}$ shunt $r_{sd.PO}$ in parallel with $R_{FB1} + R_{FB2}$ at considerably lower frequencies than the equivalent Miller pole does:

$$\left.\frac{1}{s(C_O + C_{DB.PO})}\right|_{p_O \approx \frac{1}{2\pi[r_{ds.PO} \,||\, (R_{FB1} + R_{FB2})](C_O + C_{DB.PO})}} \equiv r_{ds.PO} \,||\, (R_{FB1} + R_{FB2}) \tag{7.29}$$

Miller pole p_{EA} would then result when C_C and $C_{GD.PO}$, now that v_o is ac ground, shunt $r_{ds.NB2}$:

$$\left.\frac{1}{s(C_C + C_{GD.PO})}\right|_{p_{EA} \approx \frac{1}{2\pi r_{ds.NB2}(C_C + C_{GD.PO})} \geq \frac{f_{0dB}}{10}} \equiv r_{ds.NB2} \tag{7.30}$$

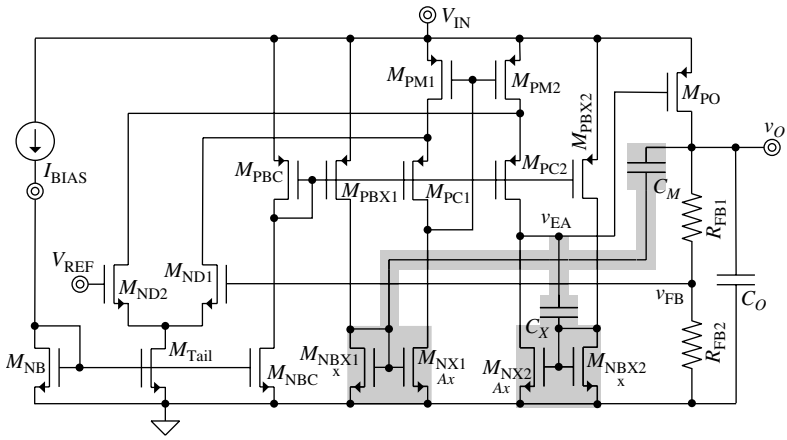
Resistor R_N would now yield the effects of a left-half-plane zero z_N when C_C's impedance shorts with respect to R_N, in other words, when $1/sC_C$ falls below R_N:

$$\left.\frac{1}{sC_C}\right|_{z_N \approx \frac{1}{2\pi R_N C_C} \approx p_{EA}} \equiv R_N \tag{7.31}$$

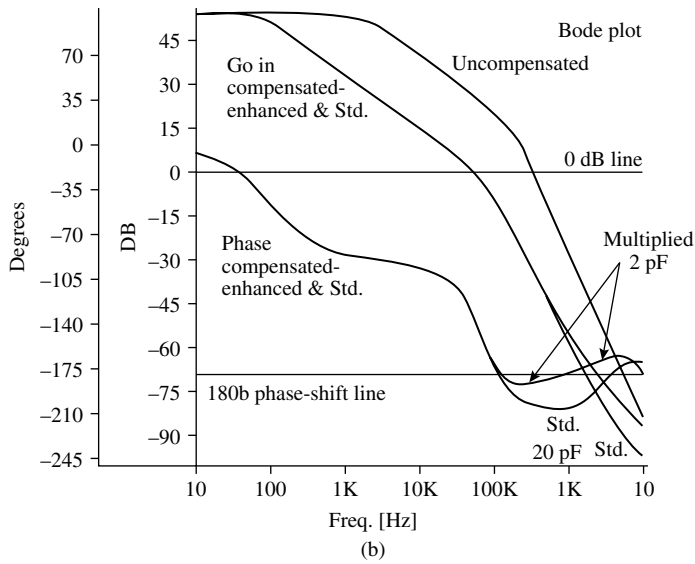
past which point the total impedance across the C_C-R_N path reduces to R_N. Under these new operating conditions, which amount to swapping p_{EA} and p_O and using z_N to cancel p_{EA} in the Bode-plot response shown in Fig. 7.3*b*, the circuit is again stable.

Worst-case stability conditions occur when C_O is at an intermediate point (e.g., 500 nF–2 µF), when p_{EA} and p_O are close. Consider, however, that not only does z_N save phase in this region but the low loop gain that results from having to accommodate load-induced variations in p_O (as discussed in Chap. 4) also helps the loop gain reach f_{0dB} without experiencing the full phase-shift impact poles p_{EA} and p_O would have otherwise had. As a result, the system is ultimately capable of remaining stable across a vast range of C_O values, albeit with varying degrees of gain and phase margins and settling-response performance. Incidentally, this role reversal between p_{EA} and p_O amounts to shifting the internally compensated strategy for its externally compensated counterpart. Also note that, with respect to power-supply rejection (PSR), Miller capacitors C_C and $C_{DG.PO}$ feed forward ac ripples in v_{IN} to v_O because shorting M_{PO}'s gate-drain terminals decreases the coupling impedance between v_{IN} and v_O to roughly $1/g_{m.PO}$, which means PSR performance suffers, underperforming the LDO and HDO circuits shown in Figs. 7.1 and 7.2.

The silicon real estate a large Miller capacitor demands from a dense system-on-chip (SoC) solution is costly and sometimes difficult to justify. Decreasing the area overhead of on-chip capacitors by actively multiplying their capacitances (as discussed in Sec. 4.6.3, Multiplying the Miller Effect) presents some risk but the area they save may nonetheless offer appealing alternatives. Figure 7.4*a* illustrates and highlights how to leverage and modify the LDO circuit presented in Fig. 7.3 to multiply the effects of capacitors C_M and C_X on the dominant low-frequency pole present at v_{EA}. The basic idea is to channel the capacitors' displacement current into amplifying current mirrors (e.g., M_{NBX1}-M_{NX1} and M_{NBX2}-M_{NX2}) and steering the multiplied currents back to v_{EA}. Mirror M_{NBX2}-M_{NX2}, for example, multiplies C_X's current i_{CX} so the total



(a)



(b)

FIGURE 7.4 Miller-multiplied low-dropout CMOS regulator (*a*) circuit and (*b*) corresponding gain-phase response across frequency.

ac current displaced from v_{EA} due to C_X is the equivalent of a capacitor C_{EQX} with $(1 + A)$ times the capacitance of C_X:

$$i_{ea} = i_{CX} + i_{NX2} = i_{CX}(1+A) \approx \left[\frac{v_{ea}}{\left(\frac{1}{sC_X}\right)}\right](1+A) \equiv \frac{v_{ea}}{\left(\frac{1}{sC_{EQX}}\right)} \qquad (7.32)$$

Similarly, but now applied to Miller capacitor C_M, mirror M_{NBX1}-M_{NX1} displaces current from v_{EA} at the equivalent rate of an A times larger C_M Miller capacitor. In other words, C_M presents a Miller-multiplied capacitance C_{EQM} at v_{EA} that is roughly equivalent to AC_M. Incidentally, C_M feeds current to v_{EA} through M_{NBX1}-M_{NX1} and M_{PM1}-M_{PM2} (instead of M_{NBX2}-M_{NX2}) to ensure its connectivity to v_{EA} remains noninverting because a conventional Miller capacitor offers no inversion from v_O to v_{EA} to maintain negative-feedback conditions across the inverting amplifier C_M encloses, which in this case is M_{PO}'s common-source gain.

Figure 7.4*b* verifies that a Miller-multiplied on-chip 2-pF capacitor achieves the same Miller objectives of a conventional 20-pF Miller capacitor. Their frequency responses differ slightly at higher frequencies because (1) the current multiplier has additional energy-shunting poles in its path (i.e., the loop is bandwidth limited) and (2) the multiplying mirror presents a series $1/g_m$ resistance that introduces what amounts to a left-half-plane zero. These nonidealities produce a peaking effect near the unity-gain frequency that C_X, which is not a Miller capacitor and benefits from having a shorter and therefore higher bandwidth path, dampens. (Incidentally, C_X is 2 pF in the results shown in Fig. 7.4*b*.) Decreasing M_{NBX1}'s $1/g_m$ resistance by, for instance, increasing the shunt-feedback loop gain across M_{NBX1}'s drain-gate terminals, that is, by introducing a high-bandwidth noninverting amplifier between M_{NBX1}'s drain and gate terminals, also dampens this effect. The biggest drawback to this technique is mirrors M_{PBX1}-M_{PBX2}, M_{NBX1}-M_{NX1}, and M_{NBX2}-M_{NX2} increase the input-referred offset of error amplifier A_{EA} because compounded mismatches in these mirrors produce a considerable offset current between M_{NX1} and M_{NX2}, which means the overall accuracy of the regulator suffers.

7.3 Self-Referencing

Accuracy across temperature is a key performance parameter, and although the regulator's input-referred offset plays a key role, reference voltage V_{REF} carries the bulk of the burden. Delegating the task to another circuit is usually prudent because other circuits in the system also rely on an accurate reference and sharing the same resource reduces overhead in silicon area, power, and possibly test time. The value of an extremely accurate reference, however, is lost in a regulator because the temperature drift of the regulator's input-referred offset counters the efforts of the reference. Dedicating a reference to the regulator, on the other hand, to correct this error is often too costly to consider, which is why reducing the offset in the first place remains the top priority.

Some applications call for self-referencing features in regulators because a global reference is not always available. Power-moding

for the sake of extending battery life, for instance, often involves powering down several functional units in a system at a time, including the reference, but not necessarily the supply. Other applications demand a self-referenced regulator simply because the only available reference is off chip, and the cost associated with allocating a pin and PCB real estate to import that reference into the IC is prohibitively high.

7.3.1 Zero-Order Approach

Temperature independence in a regulator, like dropout performance, is not always a requirement, and overdesigning a circuit for sheer accomplishment increases silicon area, power, risk, and effort unnecessarily. Irrespective of what the application demands, understanding the basics of temperature dependence is important when designing circuits to survive standard and extended commercial temperature ranges, such as 0–85°C and −40–125°C. To this end, a Taylor-series expansion is useful because it describes a parameter's dependence to temperature with respect to its zero, linear, quadratic, and higher order terms:

$$P = \sum_{A=0}^{N} K_A T^A = K_0 + K_1 T^1 + K_2 T^2 + \cdots + K_N T^N \tag{7.33}$$

where P is any parameter and K_0, K_1, K_2, and so on are its zero-, first-, second-, and higher-order coefficients. For example, the first- and higher-order coefficients of a perfectly temperature-independent reference are zero. By convention, as a result, when only the first- and second-order coefficients are zero, the parameter is more temperature dependent and said to be independent to second order. Similarly, when the first-order coefficient is nonzero and not negligibly small, irrespective of higher-order terms, the parameter is temperature independent to zero-order, which is where self-referenced regulators without temperature independent features fall.

Base-emitter (or diode), gate-source, and Zener voltages are valuable zero-order references because they produce reasonably predictable voltages across process extremes. In modern regulator applications, however, zener voltages (e.g., 5–8 V) are not as practical as their base-emitter and gate-source counterparts are because they often exceed the voltage breakdown levels of driving process technologies (e.g., 1.8–5 V). As a result, deriving a regulator output v_O, as illustrated in Fig. 7.5, from a base-emitter or gate-source voltage (i.e., Q_{NEA}'s v_{BE} or M_{NEA}'s v_{GS}) and regulating its value to, for example, $v_{BE}(R_{FB1} + R_{FB2})/R_{FB2}$ or $v_{GS}(R_{FB1} + R_{FB2})/R_{FB2}$ with a power device (i.e., Q_{NO} or M_{NO}) in shunt feedback is a popular means of self-referencing a regulator with zero-order temperature independence.

In the circuit shown, Q_{NO} and M_{NO} constitute high-dropout (HDO) power devices with relatively low output-resistance characteristics.

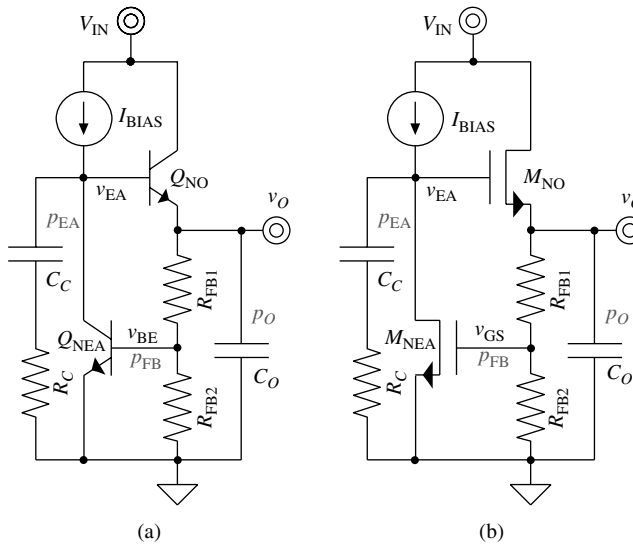


Figure 7.5 Self-referenced zero-order (*a*) BJT and (*b*) MOSFET HDO regulators.

Because v_O exhibits low resistance, output pole p_O is nondominant, but since capacitance C_O is nonetheless important and nonnegligible to supply fast load-dump currents, capacitor C_C pulls Q_{NO}'s base (and M_{NO}'s gate) pole p_{EA} to lower frequencies and ensures p_{EA} remains dominant. Feedback pole p_{FB} at v_{BE} (and v_{GS}) is not negligibly high because Q_{NEA}'s C_π (and M_{NEA}'s C_{GS}) and feedback resistances R_{FB1} and R_{FB2} are moderate (because low resistance values would otherwise increase quiescent current) so R_C introduces a left-half-plane zero to offset the combined effects of p_O and p_{FB}. Note that referencing transistors Q_{NEA} and M_{NEA} double as error amplifiers in this configuration.

7.3.2 Temperature Independence

Gate-Source Voltage

Including temperature independence increases the complexity of the circuit, although not necessarily to an extreme. Perhaps the simplest means of doing so is to find a naturally existing temperature-independent voltage or current, except such a voltage or current rarely exists in practice. Gate-source voltage v_{GS}, however, decomposes into threshold voltage V_T and drain-source saturation voltage $V_{DS(sat)}$, where the latter includes the effects of transconductance parameter K' in its denominator:

$$v_{GS} = V_T + V_{DS(sat)} = V_T + \sqrt{\frac{2i_D}{K'\left(\frac{W}{L}\right)}} \tag{7.34}$$

What is interesting about this decomposition is that V_T and K' both decrease with increasing temperatures, which means the former has a tendency to decrease v_{GS} and the latter to increase it. Since a MOSFET's bias drain current I_D and aspect ratio (W/L) affect how much the temperature drift in K' alters V_{GS}, an optimum current-density setting $I_D/(W/L)$ exists that exactly matches and opposes the effects of V_T's drift in V_{GS}. As a result, Fig. 7.5*b* also illustrates what could be a temperature-independent output, but only when $I_D/(W/L)$ is at its optimum point. In practice, process variations in V_T and K' are appreciable at maybe ±100–150 mV and ±15–25% and uncorrelated so V_{GS}'s variation drift, although reduced, remains far from zero. Nevertheless, the concept and circuit are sufficiently simple and robust to win the favor of many designers.

Bandgap Approach

Increasing temperature independence in a circuit, in a more general sense, similarly reduces to combining temperature-dependent components whose complementary effects with temperature cancel. The most practical approach is to find two voltages (or currents) whose temperature-induced variations generally oppose one another, as shown with V_{BE} or V_{CTAT} and V_{PTAT} in Fig. 7.6*a*, so that their net sum (i.e., V_{REF}) remains nearly constant with respect to temperature.

The temperature dependence of the two components that comprise the reference need not be linear (or nonlinear, for that matter) to produce a nearly constant response, just as long as they cancel each other's effects. In searching for the best candidate, looking for voltages (or currents) that do not vary significantly across process corners (from die to die, wafer to wafer, and lot to lot) is important because wide tolerances ultimately defeat the temperature-independent qualities of a supposedly accurate and predictable reference. From this point of view, the value of pn-junction diode voltage V_D (or base-emitter voltage V_{BE}) is undeniably high with respect to their gate-source counterparts (and the threshold voltages that partially define

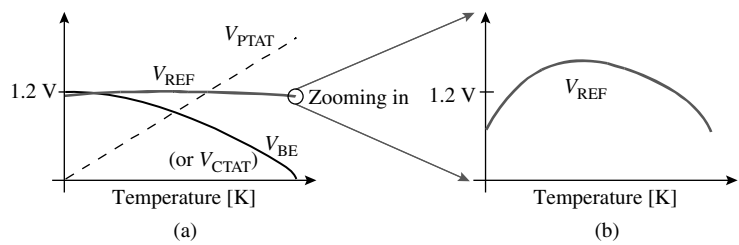


Figure 7.6 Canceling the temperature drift of a voltage that is proportional to absolute temperature (PTAT) with another that is complementary to absolute temperature (CTAT), such as diode or base-emitter voltage V_D or V_{BE}, to produce temperature-independent reference V_{REF}.

them) because the absolute variation of the former across process corners is considerably lower.

A pn-junction voltage generally decreases with increasing temperatures at approximately −2.2 mV per degree Celsius. Because its linear component for the most part overwhelms higher-order terms, diode voltages often complement voltages that are *proportional to absolute temperature* (PTAT), which is why they are said to carry *complementary-to-absolute-temperature* (CTAT) features. Ultimately, even though the compensated reference V_{REF} is mostly flat across temperature (as shown in Fig. 7.6*a*), the higher-order components in V_{BE} add the characteristic curvature shown in Fig. 7.6*b*, albeit at smaller scales.

Canceling the first-order component, as achieved by the combined effects of V_{CTAT} and V_{PTAT}, constitutes what is coined as *first-order compensation* and reducing the effects of higher-order terms is *curvature correction*, the former of which normally achieves (when trimmed) 20–100 ppm/°C of drift and the latter less than 20 ppm/°C. Even if curvature correction were to perfectly compensate higher-order components, the fillers used in the plastic package to decrease the thermal coefficient of expansion of the plastic introduce uncompensated and uncorrelated offsets in the circuit that often render curvature compensation efforts ineffectual. Ultimately, the additional circuit and trimming complexity required to improve the performance of a first-order reference by a small margin is often times difficult to justify.

The *bandgap* reference, as it is typically called, is a popular circuit that uses, at its core, a PTAT voltage V_{PTAT} to compensate the CTAT behavior inherent to diode or base-emitter voltage V_D or V_{BE}. The name bandgap arises because the zero-order temperature term of the diode voltage the reference employs is the "band-gap" voltage of silicon, which is why the diode voltage is 1.2 V at 0 K. Additionally, and perhaps more importantly, because the PTAT voltage used to compensate V_D has no zero-order term (i.e., V_{PTAT} is zero at 0 K), the reference voltage of a bandgap circuit also reduces to approximately 1.2 V, since canceling first- and higher-order terms leaves V_D's zero-order component intact. Generally, the same applies to diode- and PTAT-derived currents, except for the transresistance translation of the bandgap voltage (i.e., 1.2 V) to a resistor-defined current. Similarly, diode and PTAT fractions combine to yield references whose values fall below the bandgap mark of 1.2 V, achieving sub-bandgap values in the 0.2–1 V range. Irrespective of what ratio or derivative the circuit ultimately combines, using diode and PTAT voltages in any form to produce references follows the bandgap approach, even when applied to the core of a regulator.

Diode voltage V_D without its PTAT counterpart may be of little consequence but what is interesting about the bandgap is that V_D itself can generate its own compensating PTAT voltage. This trait results

because the exponential expression for diode current I_D includes thermal voltage V_t, which increases linearly with temperature:

$$I_D \approx I_S e^{\left(\frac{V_D}{V_t}\right)} \tag{7.35}$$

or

$$V_D \approx V_t \ln\left(\frac{I_D}{I_S}\right) \propto T^1 \tag{7.36}$$

where reverse-saturation current I_S is proportional to the cross-sectional area of the pn junction, as in the emitter area of a BJT. As a result, taking the difference between two diode voltages derived from matched, but ratioed devices (i.e., area of one is a user-defined fraction K_A of the other: A_2 is $A_1 K_A$) carrying matched and maybe ratioed currents (i.e., one current is the mirror ratio K_I of the other: I_{D1} is $I_{D2} K_I$) produces a predictable, reliable, and considerably linear PTAT voltage V_{PTAT} across temperature:

$$V_{PTAT} = \Delta V_D = V_{D1} - V_{D2} \approx V_t \ln\left(\frac{I_{D1} I_{S2}}{I_{S1} I_{D2}}\right) = V_t \ln\left(K_A K_I\right) \propto T^1 \tag{7.37}$$

The important conclusion to draw at this point is that a difference in diode or base-emitter voltages ΔV_D or ΔV_{BE} with proportionately ratioed current densities I_D/A_D produces PTAT voltage V_{PTAT}. The circuit core shown in Fig. 7.7*a*, for example, exploits this basic theorem by inserting a resistor (e.g., R_{PTAT}) in the path of what would have otherwise been two equal base-emitter voltages—the voltage across R_{PTAT} is the difference between Q_1 and Q_2's base-emitter voltages. The resulting ΔV_{BE} is PTAT only because the mirror circuit attached to Q_1 and Q_2 forces their corresponding current densities to remain constant ratios of one another. Note this same PTAT core (with additional supply- and ground-referenced mirroring transistors) is the basis for most bias current generators used today.

Although not always the case, diode-connecting one of the bipolar transistors is typical, as indicated by the dashed line in the figure, but a more general implementation of the core includes a level-shifting and maybe amplifying voltage buffer between the same collector-base terminals. For improved matching performance, the current mirror normally has a gain of 1 A/A because lower spreads in gain lead to reduced mismatches in currents. For similar reasons, only equally sized BJT segments are typically used. A ratio of eight between the matching BJTs is usually optimal because surrounding one device with eight others, as shown in Fig. 7.7*b*, is more area efficient and modular (i.e., modular or square layouts match better and are more compact) with common-centroid and two-dimensional gradient-cancellation features than just using four (or ten, or twelve) BJTs along the periphery. A ratio of two, as illustrated in Fig. 7.7*c*, is also popular for lower accuracy applications because the resulting layout is still area

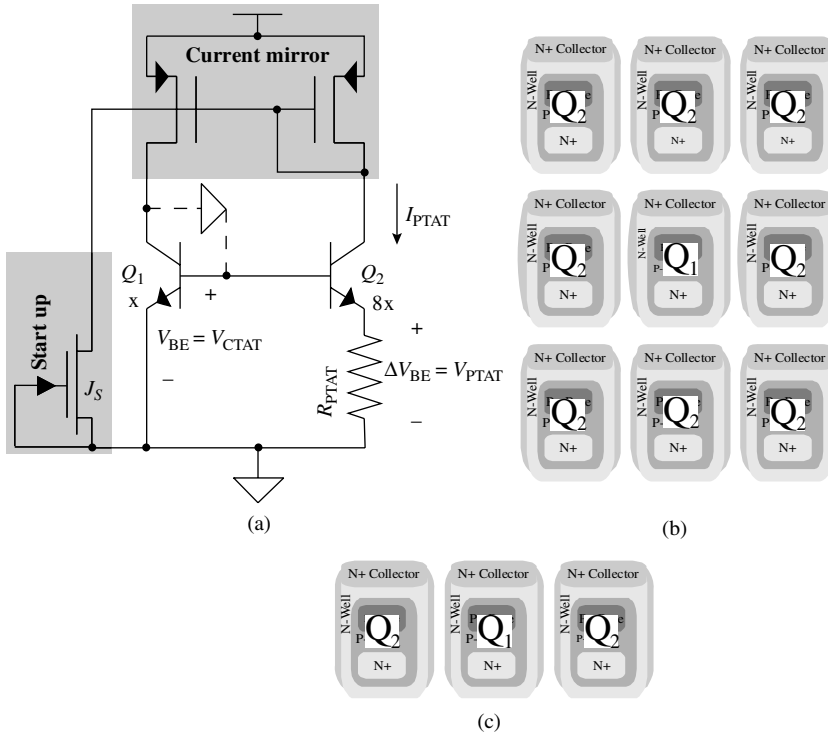


FIGURE 7.7 ΔV_{BE} PTAT generator core (*a*) circuit and (*b*) and (*c*) common BJT layout implementations.

efficient with common-centroid features, although some of its modularity and two-dimensional gradient-cancellation qualities are lost.

The circuit core, as shown in Fig. 7.7*a*, is a bi-stable circuit because transistors are just as content to carry PTAT current I_{PTAT} (or equivalently V_{PTAT}/R_{PTAT}) as they are conducting zero amps. In other words, the mirror sources ratioed currents and the ΔV_{BE} circuit ensures the currents are PTAT, unless the mirror is off, in which case the ΔV_{BE} circuit does not work. To avoid this zero-current state, a start-up circuit typically samples this current (or its voltage translation), compares it against a nominal threshold, and ensures it stays above the threshold by sourcing or sinking additional current into the circuit. Sometimes this start-up circuit is nothing more than a small and constantly flowing current into the base of the BJT pair, like from a long-channel junction field-effect transistor (JFET), as depicted by start-up JFET J_S in Fig. 7.7*a*.

Bandgap Conversion

The only difference between a zero-order diode-derived regulator and its first-order counterpart is the inclusion of a PTAT component in output v_O. Consider, for instance, how the self-referencing zero-order

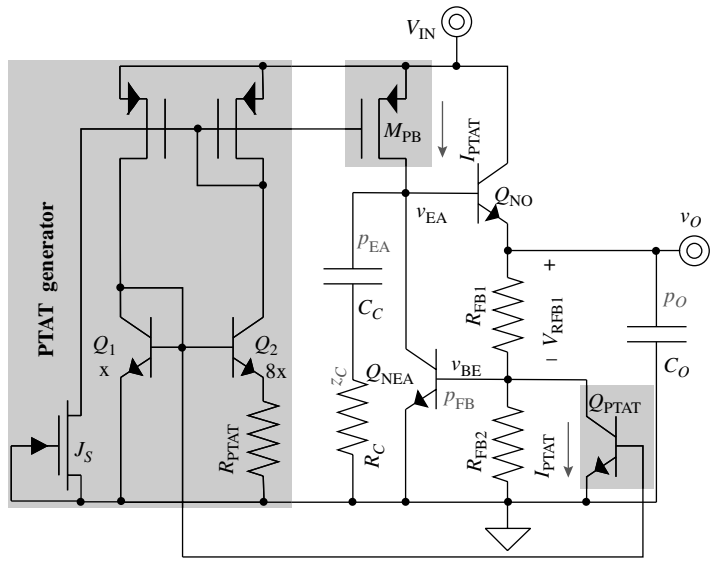


FIGURE 7.8 Inclusion of PTAT current into zero-order self-referenced HDO circuit.

regulator in Fig. 7.5*a* and the PTAT current-generator core in Fig. 7.7*a* combine to produce the first-order circuit shown in Fig. 7.8. The basic idea, with reference to the zero-order circuit in Fig. 7.5, is to incorporate PTAT voltage $\Delta V_{BE}R_{FB1}/R_{PTAT}$ into v_O by sinking a PTAT current from v_{BE} (with Q_{PTAT}), which forces a PTAT voltage component to appear in R_{FB1}'s voltage V_{RFB1}:

$$V_O = V_{BE} + \left(\frac{V_{BE}}{R_{FB2}} + I_{PTAT}\right)R_{FB1} = V_{BE} + \left(\frac{V_{BE}}{R_{FB2}} + \frac{\Delta V_{BE}}{R_{PTAT}}\right)R_{FB1}$$

$$= V_{BE}\left(\frac{R_{FB1} + R_{FB2}}{R_{FB2}}\right) + \Delta V_{BE}\left(\frac{R_{FB1}}{R_{PTAT}}\right) \tag{7.38}$$

In this case, the PTAT generator core also biases error-amplifier and referencing transistor Q_{NEA} (to conduct I_{PTAT}) via mirroring device M_{PB}.

Bandgap Integration

Another variation of the same theme is to integrate the PTAT core circuit into the feedback network of the zero-order self-referenced HDO circuit, as shown in Fig. 7.9. As before, the core pulls a PTAT current from base-emitter node v_{BE} that creates a series temperature-compensating PTAT voltage across resistor R_{P1} and in output V_O:

$$V_O = V_{BE} + I_{PTAT}R_{P1} = V_{BE} + \Delta V_{BE}\left(\frac{R_{P2}}{R_{PTAT}}\right) \tag{7.39}$$

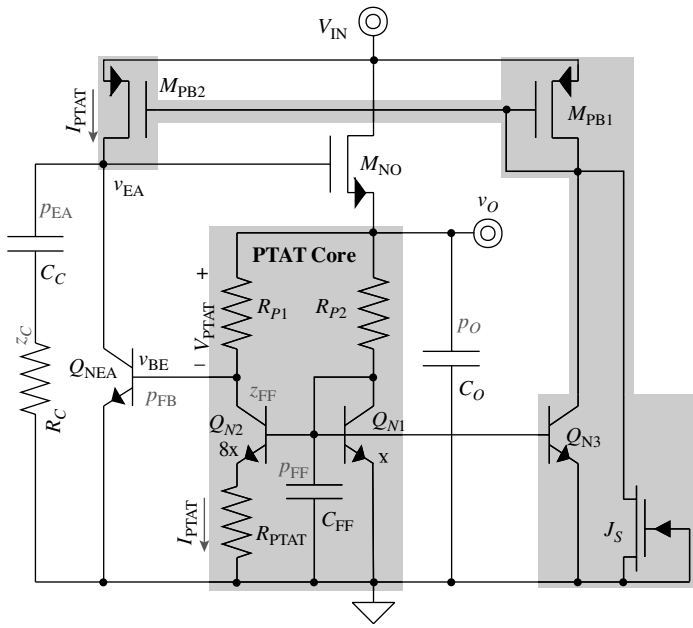


Figure 7.9 Integration of PTAT core into zero-order self-referenced HDO circuit.

In this case, resistors R_{P1} and R_{P2} in the core match and constitute a current mirror because the voltages across them equal at $v_O - v_{BE}$, where Q_{N1} and Q_{NEA}'s base-emitter voltages equal because Q_{N3}, M_{PB1}, and M_{PB2} mirror, fold, and force Q_{N1} and Q_{NEA} to carry the same current densities. N-channel start-up JFET J_S ensures v_{EA} does not drop to 0 V so the circuit does not fall into a zero-current state.

With respect to stability, C_C pulls v_{EA}'s pole p_{EA} to lower frequencies, R_C introduces left-half-plane zero z_C to mitigate the avalanching effects output pole p_O and feedback pole p_{FB} have on phase response near the unity-gain frequency, and C_O sources the load current which the regulator is not quick enough to supply during fast load dumps. PTAT-core components R_{P2}, Q_{N1}, and Q_{N2} present an out-of-phase feed-forward path to v_{BE} whose right-half-plane zero effects (i.e., z_{FF}) appear when Q_{N2}'s small-signal collector current exceeds R_{P1}'s counterpart. To push z_{FF} to higher frequencies, the degenerating impact of R_{PTAT} is applied to Q_{N2} and its resulting small-signal current, not Q_{N1}. Additionally, as frequency increases, C_{FF} shunts some of this feed-forward energy to ground, further pushing z_{FF} to higher frequencies under the guise of feed-forward pole p_{FF}.

The dual feed-forward paths in the core actually add design flexibility because reversing the polarity of the main path amounts to exchanging Q_{N1} and Q_{N2}-R_{PTAT}'s respective places in the circuit. This polarity inversion is exactly what an LDO transformation requires, as

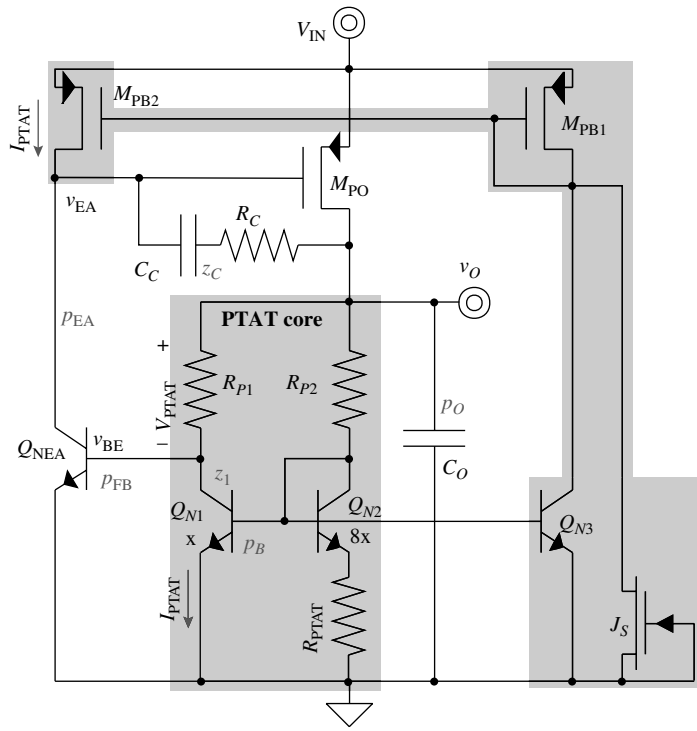


Figure 7.10 Integration of PTAT core into zero-order self-referenced LDO circuit.

shown in Fig. 7.10, because a p-type power device introduces an inversion its n-type counterpart does not. The only other variation to the circuit relates to compensation, as C_C now enjoys the benefits of Miller multiplication and R_C shifts the out-of-phase feed-forward zero across M_{PO} to the left half of the s plane. Q_{N1} and Q_{N2}'s base pole p_B is now part of the main loop, except it resides at relatively high frequencies because Q_{N2}'s diode connection's $1/g_m$ resistance is low and R_{PTAT}'s resistance is usually only moderate. Q_{N1}'s C_μ also introduces an out-of-phase feed-forward path but its resulting right-half-plane zero is also at relatively high frequencies because C_μ is small and Q_{N1}'s transconductance is high (given Q_{N1} is a BJT).

The underlying aim behind the design of the PTAT core shown in Fig. 7.7*a* is to insert R_{PTAT} in the voltage loop enclosed by the base-emitter voltages of two BJTs so that PTAT voltage difference ΔV_{BE} appears across R_{PTAT}. Where in the loop R_{PTAT} lies does not change ΔV_{BE}, as long as current densities into the transistors remain ratioed. Figure 7.11 achieves the same result by shifting R_{PTAT} out of the emitter degenerating position into the base junction of the two BJTs, keeping R_{PTAT} in the loop but now between the two bases. (Note R_{B1} and R_{B2}

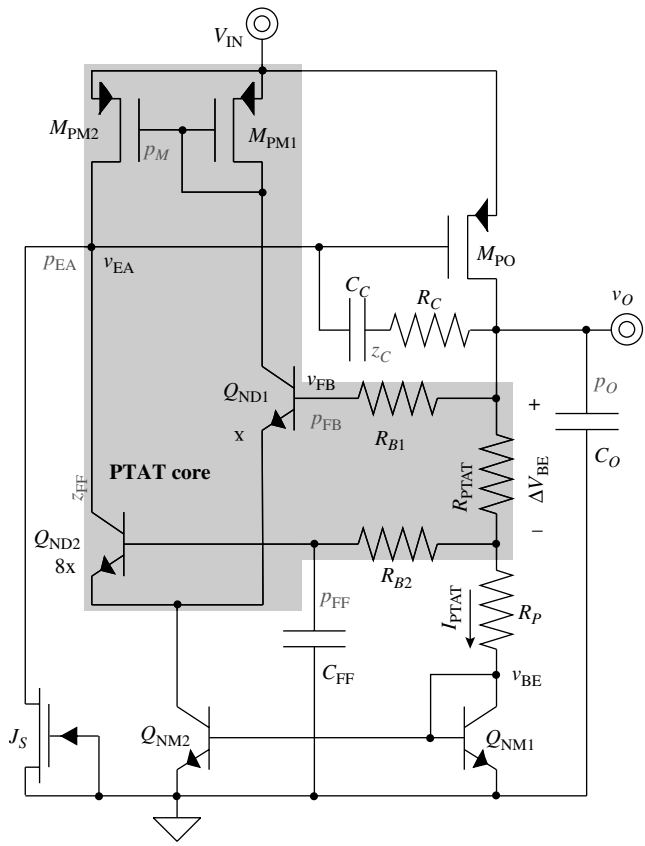


Figure 7.11 Integration of differential PTAT core into zero-order self-referenced LDO circuit.

match and introduce no offsets because their respective voltages equal.) Resistor R_p amplifies the PTAT voltage contribution in output v_O enough to offset Q_{NM1} base-emitter voltage v_{BE}'s first-order dependence and produce a bandgap-like response:

$$V_O = V_{BE} + I_{PTAT}(R_{PTAT} + R_P) = V_{BE} + \Delta V_{BE}\left(1 + \frac{R_P}{R_{PTAT}}\right) \qquad (7.40)$$

Devices M_{PO}, R_{PTAT}, and R_{B2}, with reference to the core illustrated in Fig. 7.7*a*, constitute the buffer that replaces the diode connection that would have otherwise existed around Q_{ND2}. Mirror Q_{NM1}-Q_{NM2}, like a differential tail current, ensures the total current flowing through Q_{ND1} and Q_{ND2} is PTAT and mirror M_{PM1}-M_{PM2} forces that current to split equally between the two BJTs. JFET J_S, as before, keeps the

circuit from entering the zero-current state by constantly pulling current from v_{EA} and consequently inducing M_{PO} to conduct and start the circuit, if nothing else is on.

The error amplifier in the feedback loop now reduces to differential pair Q_{ND1}-Q_{ND2} and p-type mirror load M_{PM1}-M_{PM2}, except only Q_{ND1} conducts the desired negative-feedback signal back to power device M_{PO}. R_{PTAT}, however, drops the ac value of the small signal present at v_O and reduces the out-of-phase feed-forward gain through Q_{ND2} with respect to Q_{ND1}, effectively pushing right-half-plane zero z_{FF} to higher frequencies. R_{B2}-C_{FF}'s pole p_{FF} further filters the feed-forward path to reduce its overall impact on the feedback loop. Base current produces a voltage across R_{B2} so R_{B1} is added to match that voltage and cancel any effect it would have otherwise had on the modified PTAT core. Besides amplifying the PTAT contribution in v_O, R_P also sets the headroom voltage across tail current transistor Q_{NM2} to ensure Q_{NM2}'s collector-emitter voltage remains above $V_{CE(min)}$ across the entire temperature range. C_C, R_C, and C_O, as before, pull pole p_{EA} to lower frequencies, pull zero z_C to the left half of the s plane, and source fast load currents, respectively.

7.4 Current Regulation

In power supplies, current regulation is not really an objective but a means to an end. Although currents used to bias the constituent analog and/or mixed-signal building blocks in a given system could benefit from current regulation, for example, they normally do not require it. Switching power supplies, especially boost regulators, on the other hand, often embody current regulation features to regulate inductor current and consequently transform an LC response into an RC equivalent, except these current loops do not stand on their own—voltage regulating loops normally embed them. From an analog-design perspective, perhaps the most appealing feature of current regulation is high output resistance. With respect to power supplies, for instance, testing the supply's response against "controlled" changes in load current is important so experimental setups often benefit from the use of regulated load currents. In any event, for these reasons and for the sake of completeness, the following discussion briefly addresses some important design considerations when employing current regulation.

Regulating current involves using feedback to ensure output current undergoes little to no variations when subjected to changes in output voltage, which is the natural by-product of a series-sampled negative-feedback loop. This feature of high output resistance is especially attractive in current sources and current mirrors, as variations in output current in these applications degrade accuracy. Extending the range for which this feature remains true is also important so low minimum output voltage $V_{OUT(min)}$ is another important metric to evaluate in circuits of this sort.

7.4.1 Current Sources

One way of defining and regulating a current is to use a differential transconductor (e.g., G_{EA}) or operational amplifier and a series-sampling transistor (e.g., M_{NO}) in a feedback loop, as illustrated in Fig. 7.12*a*, to set the voltage across and current through a series resistor (e.g., R_I). The virtual short circuit that results across G_{EA}'s input terminals ensures output current i_O is dependent on input reference voltage v_{REF} (i.e., i_O is v_{REF}/R_I) and independent of output voltage v_O. The loop only has two poles, where the resistance at M_{NO}'s gate (i.e., v_{EA}) is considerably higher than at its source (i.e., v_{FB}) so compensating capacitor C_C, although often not necessary, helps pull v_{EA}'s pole p_{EA} to low frequencies, ensuring p_{EA} remains dominant over p_{FB}. M_{NO}'s in-phase feed-forward capacitor C_{GS} helps maintain phase margin by introducing a left-half-plane zero z_{NO} soon after p_{FB}, rendering the circuit considerably robust with respect to stability.

The output resistance of regulated cascode transistor M_{NO} is the amplified extrapolation of its nonregulated, but still source-degenerated counterpart. With respect to feedback, the circuit's forward open-loop transconductance gain $A_{G.OL}$ or equivalently $i_O/(v_{REF} - v_{FB})$ is $G_{EA}R_{O.EA}g_{m.NO}$ (where $R_{O.EA}$ is G_{EA}'s output resistance) and feedback factor β_{FB} or equivalently v_{FB}/i_O is R_I so the closed-loop transconductance gain $A_{G.CL}$, which is i_O/v_{REF}, reduces to approximately R_I^{-1}:

$$A_{G.CL} \equiv \frac{i_O}{v_{REF}} = \frac{A_{G.OL}}{1 + A_{G.OL}\beta_{FB}} = \frac{G_{EA}R_{O.EA}g_{m.NO}}{1 + G_{EA}R_{O.EA}g_{m.NO}R_I} \approx \frac{1}{R_I} \qquad (7.41)$$

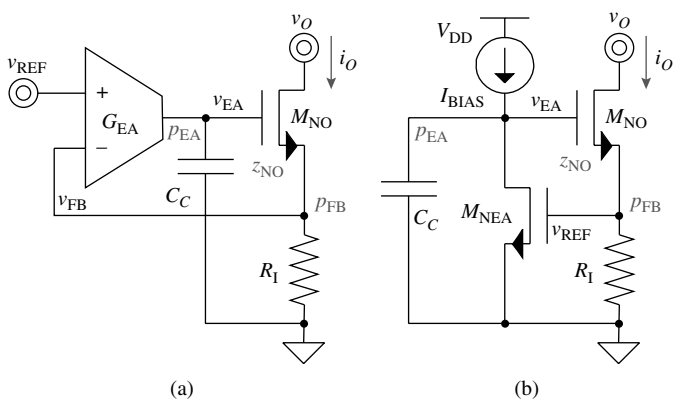


Figure 7.12 (*a*) Voltage-derived and (*b*) self-referenced regulated cascode current sources.

Series-sampled feedback increases M_{NO}'s open-loop output resistance $r_{ds.NO}$ by approximately the product of $A_{G.OL}$ (or $G_{EA}R_{O.EA}g_{m.NO}$) and β_{FB} (or R_I):

$$\begin{aligned} R_{O.CL} &= R_{O.OL}(1 + A_{G.OL}\beta_{FB}) + R_{I.FB} \\ &= r_{ds.NO}[1 + (G_{EA}R_{O.EA}g_{m.NO})R_I] + R_I \approx r_{ds.NO}(G_{EA}R_{O.EA}g_{m.NO}R_I) \end{aligned} \tag{7.42}$$

where $R_{O.CL}$ is the circuit's closed-loop output resistance and $R_{I.FB}$ the input resistance of the feedback network (i.e., R_I). Minimum output voltage $V_{OUT(min)}$ corresponds to the minimum voltage M_{NO} can sustain before entering triode (i.e., its low-gain region), which is the series combination of M_{NO}'s saturation voltage $V_{DS.NO(sat)}$ and input reference V_{REF}:

$$V_{OUT(min)} = V_{DS.NO(sat)} + V_{REF} \tag{7.43}$$

so large aspect ratios for M_{NO} and low V_{REF} values produce the best results, that is, the lowest $V_{OUT(min)}$.

Figure 7.12*b* illustrates a simplified transistor-level embodiment of the op-amp circuit shown in Fig. 7.12*a*. In this case, transistor M_{NEA} replaces transconductor G_{EA} and feedback node v_{REF} doubles as the reference, giving the circuit a self-referencing quality. The voltage across R_I is now M_{NEA}'s gate-source voltage $v_{GS.NEA}$ so the resulting output current is

$$i_O = \frac{v_{GS.NEA}}{R_I} \tag{7.44}$$

Output i_O's dc value is not regulated, as it was in the previous case, which means the circuit is more sensitive to process and temperature variations through threshold voltage V_T and transconductance parameter K'. Output current i_O is, however, regulated against variations in output voltage v_O, as it was in the previous circuit, so closed-loop output resistance $R_{O.CL}$ is similarly an amplified version of M_{NO}'s open-loop output resistance $r_{ds.NO}$:

$$\begin{aligned} R_{O.CL} &= R_{O.OL}(1 + A_{G.OL}\beta_{FB}) + R_{I.FB} \\ &= r_{ds.NO}[1 + (G_{EA}R_{O.EA}g_{m.NO})R_I] + R_I \\ &= r_{ds.NO}(1 + g_{m.NEA}r_{ds.NEA}g_{m.NO}R_I) + R_I \\ &\approx r_{ds.NO}(g_{m.NEA}r_{ds.NEA}g_{m.NO}R_I) \end{aligned} \tag{7.45}$$

The minimum output voltage for this circuit is, as before,

$$V_{O(min)} = V_{DS.NO(sat)} + V_{REF} = V_{DS.NO(sat)} + V_{GS.NEA} \tag{7.46}$$

except reference V_{REF} is necessarily a gate-source voltage above ground (e.g., 0.9–1.5 V) so $V_{OUT(min)}$ is potentially higher than its op-amp counterpart.

7.4.2 Current Mirrors

Current regulators, like their voltage regulating counterparts (but with respect to current), often rely on existing current references to derive their output, ultimately performing the function of a current mirror, but with regulating features. From this viewpoint, adding regulated cascode transistor M_{NO} to a simple current-mirror output, as shown in Fig. 7.13*a*, achieves the objectives sought in a current regulator. Output current i_O, in this case, is now a linear translation of input current i_{IN}

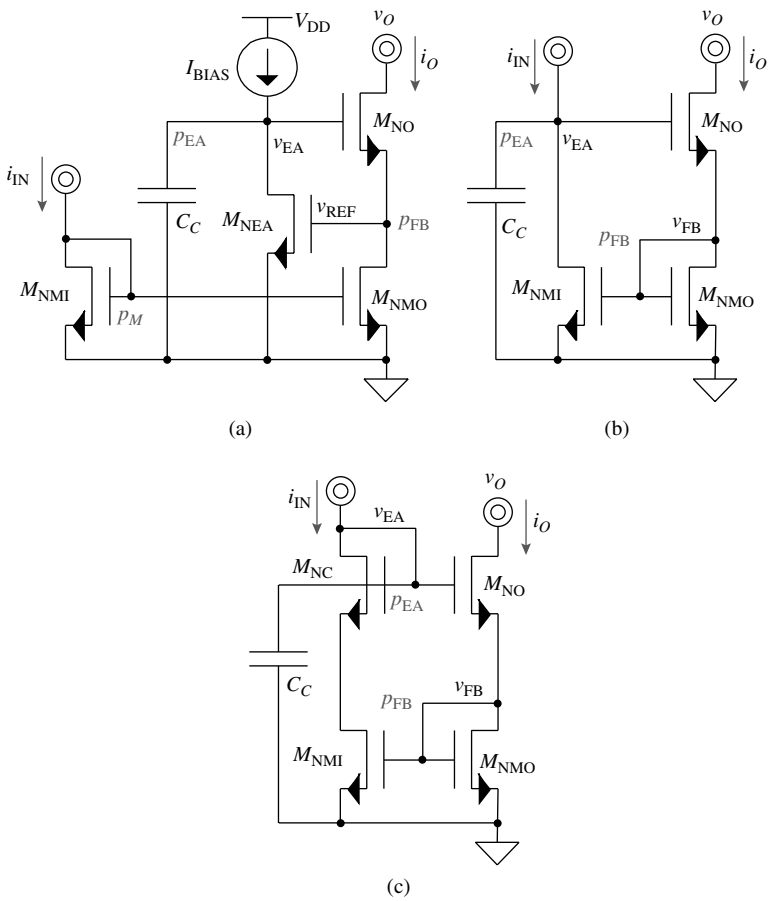


Figure 7.13 Regulated cascode current-mirror evolution from its regulated cascode current-source predecessor.

with little to no Early-voltage errors because both drain-source voltages are a gate-source voltage above ground:

$$i_O = i_{\text{IN}}\left(\frac{1+V_{\text{DS}.M1}}{1+V_{\text{DS}.M2}}\right) = i_{\text{IN}}\left(\frac{1+V_{\text{GS}.M1}}{1+V_{\text{GS.NEA}}}\right) \approx i_{\text{IN}} \tag{7.47}$$

As with the circuit shown in Fig. 7.12*b*, however, i_O is only regulated against variations in output voltage, not the matching qualities of the mirror. As such, closed-loop output resistance $R_{\text{O.CL}}$ is the same amplified version of open-loop cascode resistance $r_{\text{ds.NO}}$ $R_{\text{O.CL}}$ was in Fig. 7.12*b*, except β_{FB} and feedback input resistance $R_{I.\text{FB}}$ are now $r_{\text{ds.NMO}}$:

$$\begin{aligned} R_{\text{O.CL}} &= R_{\text{O.OL}}(1+A_{\text{G.OL}}\beta_{\text{FB}}) + R_{I.\text{FB}} \\ &= r_{\text{ds.NO}}[1+(G_{\text{EA}}R_{\text{O.EA}}g_{m.\text{NO}})r_{\text{ds.NMO}}] + r_{\text{ds.NMO}} \\ &= r_{\text{ds.NO}}(1+g_{m.\text{NEA}}r_{\text{ds.NEA}}g_{m.\text{NO}}r_{\text{ds.NMO}}) + r_{\text{ds.NMO}} \\ &\approx r_{\text{ds.NO}}(g_{m.\text{NEA}}r_{\text{ds.NEA}}g_{m.\text{NO}}r_{\text{ds.NMO}}) \end{aligned} \tag{7.48}$$

In the same vein, because the same regulating loop exists, minimum output voltage $V_{\text{OUT(min)}}$ remains the same at $V_{\text{DS.NO(sat)}} + V_{\text{GS.NEA}}$.

Figure 7.13*b* illustrates how to integrate the mirror into the feedback loop to reduce silicon real estate and power consumption, resulting in a compact current-mixed, current-sampled feedback circuit. As it pertains to feedback, forward open-loop current gain $A_{I.\text{OL}}$ or equivalently $i_O/(i_{\text{IN}} - i_{\text{NMI}})$ is now the product of current-voltage converter $r_{\text{ds.NMI}}$ and degenerated transconductance $g_{m.\text{NO}}/(1 + g_{m.\text{NO}}/g_{m.\text{NMO}})$. Feedback factor β_{FB} (or i_{NMI}/i_O) is M_{NMI}-M_{NMO} mirror ratio $g_{m.\text{NMI}}/g_{m.\text{NMO}}$ or $(W/L)_{\text{NMI}}/(W/L)_{\text{NMO}}$, which means closed-loop current gain $A_{I.\text{CL}}$ reduces to β_{FB}^{-1} or $(W/L)_{\text{NMO}}/(W/L)_{\text{NMI}}$:

$$\begin{aligned} A_{I.\text{CL}} \equiv \frac{i_O}{i_{\text{IN}}} &= \frac{A_{I.\text{OL}}}{1+A_{I.\text{OL}}\beta_{\text{FB}}} = \frac{\dfrac{r_{\text{ds.NMI}}g_{m.\text{NO}}}{\left(1+\dfrac{g_{m.\text{NO}}}{g_{m.\text{NMO}}}\right)}}{1+\dfrac{r_{\text{ds.NMI}}g_{m.\text{NO}}}{\left(1+\dfrac{g_{m.\text{NO}}}{g_{m.\text{NMO}}}\right)}\left(\dfrac{g_{m.\text{NMI}}}{g_{m.\text{NMO}}}\right)} \\ &\approx \frac{g_{m.\text{NMO}}}{g_{m.\text{NMI}}} = \frac{\left(\dfrac{W}{L}\right)_{\text{NMO}}}{\left(\dfrac{W}{L}\right)_{\text{NMI}}} \end{aligned} \tag{7.49}$$

Series-sampled feedback increases open-loop output resistance $R_{O.OL}$ to closed-loop resistance $R_{O.CL}$

$$\begin{aligned} R_{O.CL} &= R_{O.OL}(1 + A_{I.OL}\beta_{FB}) + R_{I.FB} \\ &= r_{ds.NO}\left[1 + \left(\frac{r_{ds.NMI}g_{m.NO}}{1 + \dfrac{g_{m.NO}}{g_{m.NMO}}}\right)\left(\frac{g_{m.NMI}}{g_{m.NMO}}\right)\right] + \frac{1}{g_{m.NMO}} \\ &\approx \frac{r_{ds.NO}r_{ds.NMI}g_{m.NO}}{1 + \dfrac{g_{m.NO}}{g_{m.NMO}}} \approx \frac{r_{ds.NO}r_{ds.NMI}g_{m.NO}}{2} \end{aligned} \tag{7.50}$$

which is lower than the previous circuits' because R_I is now replaced with $1/g_{m.NMO}$. The cost of simplicity in this case is accuracy because the drain-source voltages of mirror M_{NMI}-M_{NMO} do not equal:

$$i_O = i_{IN}\left(\frac{1 + V_{DS.NMI}}{1 + V_{DS.NMO}}\right) = i_{IN}\left[\frac{1 + (V_{GS.NMO} + V_{GS.NO})}{1 + V_{GS.NMO}}\right] \tag{7.51}$$

which is why level-shifting cascode transistor M_{NC} is included in Fig. 7.13*c*, to ensure these two voltages equal:

$$\begin{aligned} i_O &= i_{IN}\left(\frac{1 + V_{DS.NMI}}{1 + V_{DS.NMO}}\right) = i_{IN}\left[\frac{1 + (V_{GS.NMO} + V_{GS.NO} - V_{GS.NC})}{1 + V_{GS.NMO}}\right] \\ &\approx i_{IN}\left(\frac{1 + V_{GS.NMO}}{1 + V_{GS.NMO}}\right) = i_{IN} \end{aligned} \tag{7.52}$$

Note minimum output voltage $V_{OUT(min)}$ for all three circuit versions is similar to that of the regulated current source shown in Fig. 7.12*b* at $V_{DS(sat)} + V_{GS}$.

Other important metrics in a regulated current mirror include closed-loop input resistance $R_{I.CL}$ and minimum input voltage $V_{IN(min)}$. The closed-loop input resistance of the nonintegrated mirror is simply M_{NMI}'s diode-connected resistance $1/g_{m.NMI}$. Integrating the mirror into the loop forces the circuit to shunt-mix i_{IN} so its input resistance is still similarly low at roughly $2/g_{m.NO}$:

$$R_{I.CL} = \frac{R_{I.OL} \parallel R_{O.FB}}{1 + A_{I.OL}\beta_{FB}} = \frac{r_{ds.NMI}}{1 + \left(\dfrac{r_{ds.NMI}g_{m.NO}}{1 + \dfrac{g_{m.NO}}{g_{m.NMO}}}\right)\left(\dfrac{g_{m.NMI}}{g_{m.NMO}}\right)} \approx \frac{2}{g_{m.NO}} \tag{7.53}$$

	Basic Mirror w/R_β	Low-Voltage Cascoded Mirror	Regulated Cascode with a Nonintegrated Mirror	Regulated Cascode with Integrated Mirror and Level-Shifting M_{NC}
$V_{IN(min)}$	V_{GS}	V_{GS}	V_{GS}	$2V_{GS}$
$V_{OUT(min)}$	$V_{DS(sat)}$	$2V_{DS(sat)}$	$V_{DS(sat)} + V_{GS}$	$V_{DS(sat)} + V_{GS}$
Accuracy	λ error	–	–	–
R_{IN}	$1/g_m$	$1/g_m$	$1/g_m$	$2/g_m$
R_{OUT}	r_{ds}	$r_{ds}^2 g_m$	$r_{ds}^3 g_m^2$	$0.5 r_{ds}^2 g_m$
Notes	Moderate R_{OUT} and low accuracy	High R_{OUT} and requires biasing	Highest R_{OUT}, requires biasing, and high $V_{OUT(min)}$	High R_{OUT} and high $V_{OUT(min)}$

TABLE 7.1 Mirror Summary

With respect to $V_{IN(min)}$, the first circuit of Fig. 7.13*a*, where the mirror is outside the loop, enjoys a lower minimum input voltage at $V_{GS.NMI}$, which is approximately half of its more integrated counterparts—$V_{IN(min)}$ for the mirror-integrated loops is $V_{GS.NMO} + V_{GS.NO}$.

Summary

In view of the foregoing current-mirror discussion, a comparison of the performance of the regulated current mirrors presented in this section against its more basic predecessors discussed in Chap. 3, summarized in Table 3.2, and now revisited in Table 7.1 merits inspection. The basic benefit of the regulated cascode circuit is to amplify the effects of its nonregulated cascode counterpart, except integrating the mirror into the loop decreases the gain back to its nonregulated levels. Including a regulating loop around the cascode transistor also tends to increase $V_{OUT(min)}$, and decreasing $V_{OUT(min)}$ below its natural level complicates the circuit, sacrificing both power and accuracy performance. Because regulating the cascode transistor ultimately increases quiescent power dissipation and $V_{OUT(min)}$, doing so for the sake of a higher R_{OUT} is only worthwhile when absolutely necessary. The basic cascode circuit is often more practical because it offers a more balanced cost-performance tradeoff, and even then, the demand for high output resistance is the only justifiable reason for using it.

7.5 Low-Current and Low-Dropout Enhancements

Extended operational life is undoubtedly a driving demand for a growing number of portable, battery-powered devices such as self-powered wireless microsensors, personal digital assistants (PDAs),

cellular phones, palm pilots, biomedical implants, and so on. Aside from requiring a lower dropout voltage, extending operational life also equates to decreasing quiescent-current flow, which reduces load range, increases response time (i.e., lowers bandwidth), worsens dc accuracy, and deteriorates supply-ripple rejection. Not surprisingly, improving these basic performance parameters is the subject of much attention in the industry, especially within the context of total on-chip integration. The following subsections therefore explore ways of enhancing the LDO regulator circuit to better address these specific areas.

7.5.1 Power Switch

Much of the needs of the regulator emanate from the power pass device because it not only sets the driving load range (or equivalently, the dropout voltage) of the circuit but also presents a considerably large parasitic capacitance that necessarily slows the response of the feedback loop and its ability to respond to rapidly changing loads. Increasing the driving range of the regulator without slowing the loop amounts to boosting the gate drive of the power switch without increase its size. Two ways of achieving this goal under similar supply constraints are to (1) reduce the effective threshold voltage of the device during high load-current conditions and (2) add a supplementary out-of-the-loop "slave" transistor that sources a significant fraction of the total load, augmenting the sourcing power of the regulator without enlarging the power device.

Bulk Boost

Although a low threshold-voltage device can source higher currents, it also leaks more current (especially at high temperatures) during stand-by conditions, which drains the battery and decreases its single-charge life. Forward biasing a PMOS transistor's source-bulk pn junction "on demand," on the other hand, decreases the threshold voltage of the device (e.g., $|V_{TP}|$) only when needed, retaining low-leakage characteristics at light loads and increasing overdrive (e.g., v_{SGT}, or equivalently, $v_{SG} - |V_{TP}|$) during heavy loading conditions. The on-demand feature is achieved by sensing output transistor M_{PO}'s drain current i_{DO} with a substantially smaller sense mirror transistor M_{PS}, as shown in Fig. 7.14, and forward biasing the bulk only when M_{PS} drain current i_{DS} increases, that is to say, when a mirrored version of i_{DS} induces a voltage across bulk-driving resistor R_{B1}:

$$v_{SGT} = v_{SG} - \left|V_{TP}(V_{SB})\right| \propto i_{DS}R_{B1} \propto i_{DO} \tag{7.54}$$

Note the 1:1000 mirror-ratio gain between M_{PO} and M_{PS} is only an illustrative example, as this ratio can be higher or lower, depending on the targeted application.

Driving the bulk with a gate-derived signal (i.e., v_{BUF}) also constitutes an in-phase, feed-forward ac signal path in the feedback loop of

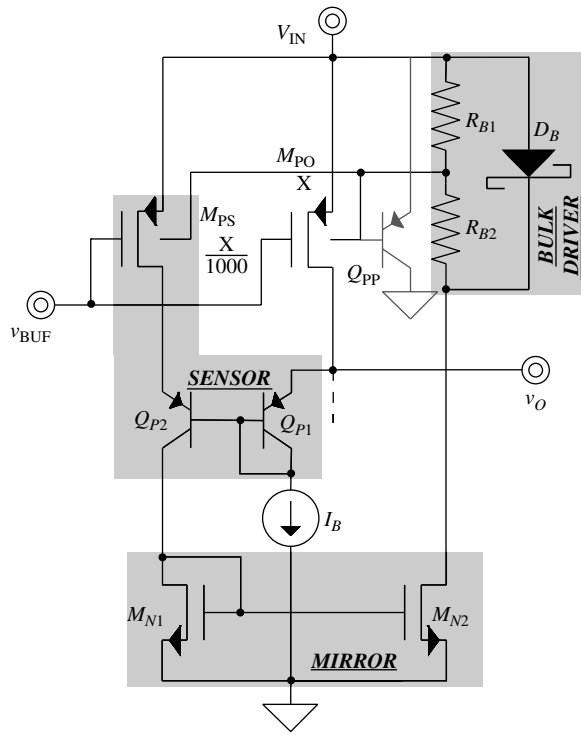


FIGURE 7.14 Bulk-driver circuit for a low-dropout PMOS power switch.

the regulator. The feed-forward gain, however, normally remains well below its gate-driven counterpart because bulk-source transconductance g_{mb} is inherently lower than g_m and diode D_B from the bulk-driver circuit loads and consequently reduces the feed-forward gain, which means a left-half-plane zero does not typically result. There is also a negative-feedback loop from output v_O through sensor Q_{P1}-Q_{P2} and mirror M_{N1}-M_{N2} back to M_{PO}'s bulk, except Q_{P2} is degenerated with M_{PS} and its gain is therefore substantially lower than the gain of the regulator's main (outer) negative-feedback loop.

Functionally, as also discussed in the buffer section of Chap. 6, bias current sensor Q_{P1}-Q_{P2} ensures M_{PS} and M_{PO}'s drain-source voltages roughly equal so their mirroring ratio remains unchanged through dropout conditions, when M_{PO} enters the triode region. Mirror M_{N1}-M_{N2} reproduces M_{PS} sense current i_{DS} and pulls it from the bulk-driver circuit. A voltage-divided (via R_{B1} and R_{B2}) Schottky-diode (i.e., D_B) voltage then limits M_{PO}'s forward-bias voltage well below the emitter-base voltage required to induce the parasitic pnp BJT attached to M_{PO}'s v_{IN} terminal (i.e., Q_{PP}) to channel considerable current into the substrate. Ultimately, as illustrated in Fig. 7.15*a*, increasing the

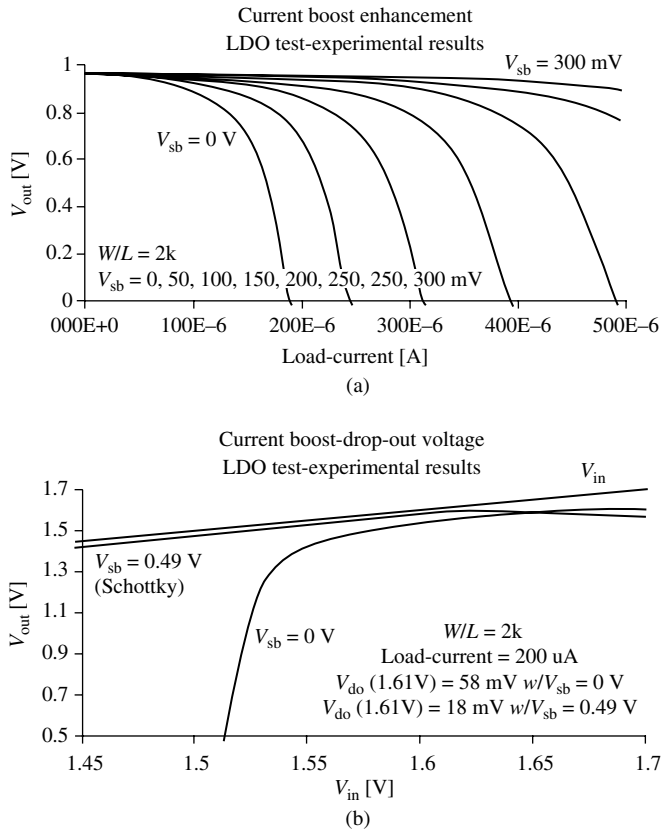


Figure 7.15 (*a*) Extending load-current range and (*b*) reducing dropout voltage by forward biasing the source-bulk junction of power PMOS transistor M_{PO}.

source-bulk voltage in this way extended the load-current range of a 2000/1 PMOS power device with a 1.2-V supply and a 0.9-V threshold voltage from 20 μA to roughly 0.5 mA. Similarly, as demonstrated in Fig. 7.15*b*, applying a 490-mV source-bulk voltage reduced the 200-μA dropout voltage from 58 to 18 mV.

Master-Slave Power Transistors

Another way of augmenting the driving capabilities of the LDO regulator without increasing the size of power switch M_{PO} in the control loop is by adding an auxiliary power device (e.g., M_{PX}) outside the loop with its control derived from M_{PO}. The basic idea is to introduce a "slave" transistor whose control is unobtrusively derived from M_{PO} and have it source additional current into v_O. Slave transistor M_{PX} therefore increases the total output current of the regulator without presenting additional capacitance to the buffer (i.e., to v_{BUF}) in the main feedback path, in other words, without slowing the main feedback path and its response time to fast load dumps.

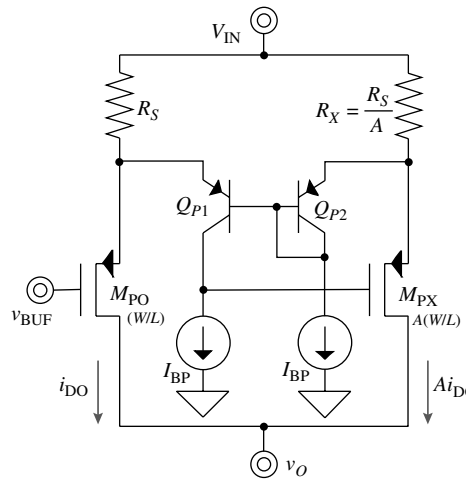


FIGURE 7.16
Linear master-slave power PMOS transistor circuit.

Linear Slave Figure 7.16 illustrates a linear embodiment of the foregoing master-slave approach. In this case, M_{PX} (through series-mixed negative feedback) sources a well-controlled and proportionately larger linear translation of M_{PO}'s drain current i_{DO}. Operationally, base-coupled differential pair Q_{P1}-Q_{P2} superimposes the ohmic voltage translation of i_{DO} across series resistor R_S (i.e., V_{RS} is $i_{DO}R_S$) on sourcing auxiliary resistor R_X (i.e., V_{RX} is also $i_{DO}R_S$). Because R_X is A times smaller than R_S and its current is consequently A times larger (i.e., i_X is Ai_S), the loop gain across the negative-feedback path through and around Q_{P2} and M_{PX} supplies whatever gate voltage is necessary for M_{PX} to source amplified current Ai_{DO}.

Although the voltages across series resistors R_S and R_X necessarily increase the effective dropout voltage of the regulator with respect to the silicon real estate used, their impact need not be excessive, especially when using metallic resistors, which present low resistances. This nominal loss in silicon efficiency, however, does not negate the benefits of increased bandwidth, since now a larger device (i.e., M_{PO}-M_{PX} composite) only presents the capacitance of a smaller transistor (i.e., M_{PO}). The elegance of the circuit, in fact, outside of its simplicity, rests on how unobtrusive the circuit is to the main feedback path of the regulator. The slave-feedback loop is even self-compensating because only one low-frequency pole exists, and that is at the gate of M_{PX}.

Noninear Slave The linear relationship between auxiliary current i_{DX} and loop current i_{DO} need not be linear for the master-slave approach to work, as long as the main loop continues to regulate i_{DO} about whatever point i_{DX} establishes; in fact, i_{DX} need not even have an ac component. In essence, supplementary power switch M_{PX} can shift the biasing current flowing through main power transistor M_{PO} by sourcing a relatively large fraction of the total steady-state load without presenting extra capacitance to the driving buffer at v_{BUF}. Consider

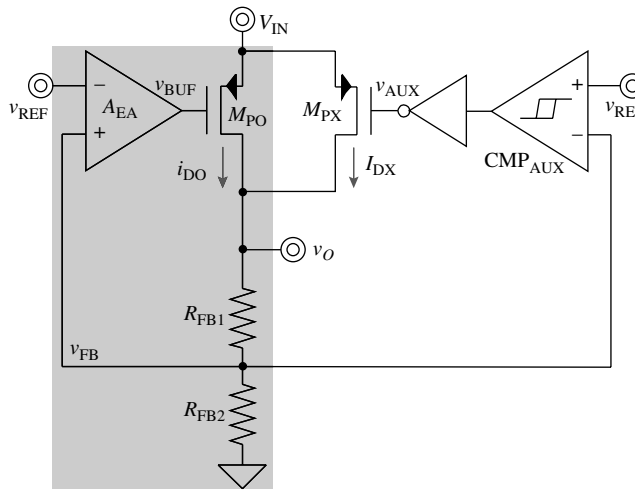


Figure 7.17 Nonlinear master-slave power PMOS transistor loop.

the case depicted in Fig. 7.17. Once load current I_L increases above a certain threshold (beyond which v_O would otherwise decrease below an acceptable window limit of v_{REF}), M_{PX} engages and sources dc current I_{DX} into v_O about which M_{PO} can now regulate, supplying only the remaining current difference (i.e., i_{DO} is $I_L - I_{DX}$).

As in the linear case, the advantage of this circuit is that the regulator can now source the current of two power transistors (i.e., M_{PO} and M_{PX}) at the speed and bandwidth of one (i.e., M_{PO}). The auxiliary path does present another negative-feedback path to v_O but its ac gain is close to zero because hysteretic comparator CMP_{AUX} latches M_{PX} on or off, not allowing M_{PX} to change until the load experiences another large-signal variation and v_O manages to reach CMP_{AUX}'s other window limit. As a result, the circuit responds quickly (at the speed of M_{PO}) in the presence of load dumps that do not require M_{PX} to switch state. Under extreme conditions, though, when load dumps exceed M_{PO}'s working range, M_{PX}'s response and transition time slow the circuit, which is why CMP_{AUX} and its driving buffer must be relatively quick. Accelerating M_{PX}'s transition is especially important in portable applications where, in an effort to save energy and extend battery life, large sections of the system may completely power up or shut down, presenting in the process severe load dumps to the regulating supplies.

7.5.2 Buffer

Bulk Feedback

Driving the base or gate of a large power device at high speed with little to no current is one of the most challenging aspects of designing an LDO regulator. To this end, the basic aim of the feedback

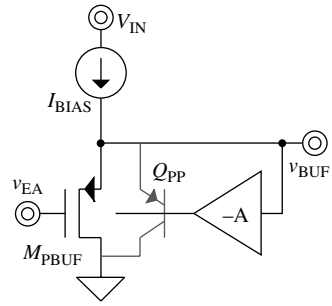


FIGURE 7.18 Bulk feedback in source following buffers.

loops presented in the buffer section of Chap. 6 was to reduce the driving impedance of emitter- and source-following buffers while dynamically adjusting (on demand) their bias point to optimally accommodate various load levels, in some instances using positive feedback to further accelerate the response. Applying the feedback signal through the bulk, in the case of bulk-isolated MOS source followers, as generally illustrated in Fig. 7.18 with PMOS buffer M_{PBUF}, is perhaps an efficient means of decreasing the output impedance without degrading the swinging limits of the buffer or considerably increasing the number of transistors in the circuit. In fact, allowing the source-bulk pn junction to forward bias decreases M_{PBUF}'s threshold voltage $|V_{TP}|$ and extends v_{BUF}'s lower swing limit closer to ground, and letting the parasitic vertical BJT attached to M_{PBUF}'s source (i.e., Q_{PP}) enter its forward-active region helps M_{PBUF} sink even more current from the gate of the power PMOS.

Speed-Up Transistor

Perhaps a more efficient means of improving the transient response of the buffer is to include a speed-up transistor that conducts and channels current to the base or gate of the large power device only during slewing conditions (on demand), when the buffer needs it most. Speed-up transistor M_{PX} in Fig. 7.19*a* does just this: stay off in

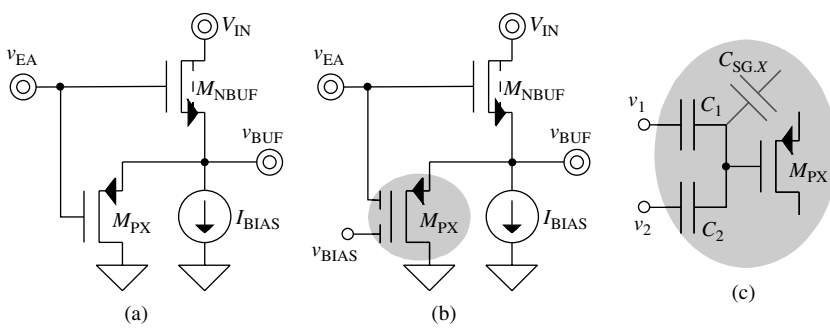


FIGURE 7.19 (*a*) Slew-rate speed-up transistor M_{PX} and (*b*) and (*c*) a capacitively coupled embodiment.

steady state, when v_{EA} is above v_{BUF}, and engage and pull additional current from v_{BUF} during large positive load dumps, when the load suddenly increases to such an extent that v_{EA} momentarily falls below v_{BUF} by at least a couple of threshold voltages. Although v_{EA} decreases and source-following buffer M_{NBUF} shuts off in response to positive load dumps at the bandwidth of the loop, the large parasitic capacitance present at v_{BUF} slews with bias current I_{BIAS}, undesirably extending the response time of the power p-type device. During severe slew-rate conditions, however, when v_{BUF} is not fast enough to decrease with v_{EA} and lags it by a sufficiently large voltage (e.g., M_{PX}'s threshold voltage $|V_{TP.X}|$), M_{PX} conducts and pulls additional current from v_{BUF}, the problematic node, accelerating the transition and the overall response time of the circuit.

The only limitation of the speed-up device is the same feature that keeps it off in steady state: its threshold voltage (i.e., $V_{TP.X}$) and that of the buffer transistor (i.e., $V_{TN.BUF}$). The problem is v_{EA} (which is normally a gate-source voltage above v_{BUF}) must traverse through at least two threshold voltages (to a $|V_{TP}|$ below v_{BUF}) before engaging M_{PX}. This means that, for M_{PX} to help, the load dump must be large and fast enough to induce error amplifier A_{EA} to swing v_{EA} two threshold voltages (i.e., $|V_{TP.X}| + V_{TN.BUF}$), at least. Replacing the buffer transistor with a natural NFET decreases this voltage by one threshold voltage, given $V_{TN.BUF}$ is near 0 V, but considering bulk effect, v_{EA} must nonetheless swing approximately 200–400 mV beyond one $|V_{TP.X}|$ for M_{PX} to conduct current. Further reductions in this voltage excursion call for lower $|V_{TP}|$ values, which is where more expensive process technologies whose mask-set composition accommodates such devices claim superiority.

Another way of decreasing $|V_{TP}|$ is through the bulk effect, by forward biasing M_{PX}'s source-bulk terminals. The drawback to this approach is ensuring the forward-bias voltage is nowhere near the voltage necessary to induce the parasitic vertical BJT attached to M_{PX}'s source to conduct ground current, which would otherwise dissipate unnecessary quiescent power. Pre-biasing M_{PX}'s source-gate voltage $v_{SG.X}$ to half its threshold-voltage point and subsequently coupling v_{EA} into its gate capacitively, as generally shown in Fig. 7.19*b* and *c*, achieves a similar result without the danger of inadvertently engaging a parasitic BJT. The idea is to precharge M_{PX}'s gate-source capacitance $C_{SG.X}$ with either a switched-capacitor network or a floating-gate input, where the gate voltage in the latter is a divider combination of its inputs:

$$v_{G.X} = \frac{v_1 Z_{C2}}{Z_{C1} + Z_{C2}} + \frac{v_2 Z_{C1}}{Z_{C1} + Z_{C2}} = \frac{v_1 C_1}{C_1 + C_2} + \frac{v_2 C_2}{C_1 + C_2} \tag{7.55}$$

and capacitors C_1 and C_2 are considerably larger than $C_{SG.X}$. The challenges with the floating-gate tactic are (1) generating a v_{BIAS} that consistently

keeps $v_{SG.X}$ at roughly $0.5 \mid V_{TP.X} \mid$ in steady state and (2) ensuring the initial charge on the gate (before applying v_{EA} and v_{BIAS}) is zero.

7.5.3 Error Amplifier

One of the challenges associated with feedback error amplifier A_{EA} is the impact limited gain has on accuracy (i.e., load and line regulation) and power-supply rejection. The problem is substantially low quiescent currents restrain parasitic poles to relatively low frequencies because the $1/g_m$ resistances they define rarely fall below 20–30 kΩ. Consider, for example, $1/g_m$ for a 5/1-µm/µm MOSFET running 0.5 µA with a transconductance parameter of 100 $\mu A/V^2$ is roughly 45 kΩ and the pole $1/g_m$ produces when confronted with 0.5 pF is $g_m/2\pi C$ or approximately 7 MHz, which means unity-gain frequency f_{0dB} must necessarily fall below 1 MHz. As a result, A_{EA}'s gain in the presence of in-band equivalent-series-resistor (ESR) zero z_{ESR} and its accompanying bypass pole p_B, as graphically demonstrated by the basic system's response in Fig. 7.20, must not exceed 40–50 dB to ensure f_{0dB} remains well below the parasitic-pole region (i.e., well below 5–10 MHz).

Another way to view the problem is through the impact z_{ESR} and p_B have on the falling rate of the loop gain. Because the zero precedes the pole, the pair, as whole, decelerates the loop gain's fall with respect to frequency, requiring a wider band of frequencies for the gain to reach 0 dB. One method of increasing this falling rate (and increasing the allowable low-frequency gain when constrained to a limited f_{0dB}) without compromising the stability of the system is to introduce a pole-zero pair well below f_{0dB}, as depicted by p_X and z_X in Fig. 7.20. The idea is for p_X to accelerate the fall and z_X to recover the phase lost to p_X before reaching f_{0dB}. The main challenge with this approach is z_{ESR} and p_B are not well controlled, which means f_{0dB} varies significantly across temperature, operating conditions, and process corners and sufficient design margin must therefore exist to accommodate all possible values of f_{0dB}.

Considering power is a precious commodity, leveraging the circuit that already exists to introduce the aforementioned pole-zero

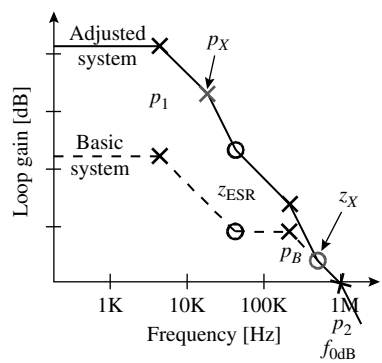


FIGURE 7.20
The effect of introducing in-band pole-zero pair p_X-z_X into the system.

combination is ideal. In an externally compensated LDO regulator, for instance, A_{EA}'s output normally constitutes a parasitic pole in the system (i.e., p_A or p_2 in Fig. 7.20) that could easily pose as the pole sought (i.e., p_X) because the resistance at v_{EA} is naturally high (at R_{EA}). Trailing a zero past p_A amounts to feed-forwarding in-phase ac signals to v_{EA} or around v_{EA} to v_O, or shaping v_{EA}'s impedance.

Figure 7.21 illustrates two embodiments of what might deceptively look like feed-forward zeros but in practice are impedance-shaping zeros. Although zero capacitor C_Z feeds forward in-phase ac signals from differential-pair node v_D and around cascode Q_C to v_{EA} in Fig. 7.21*a* (and from mirror node v_M and around mirror M_{PM2} to v_{EA} in Fig. 7.21*b*), C_Z's displacement current does not usually supersede its counterpart (i.e., Q_C's collector and M_{PM2}'s drain current) so no zero results. As it happens, Q_C and M_{PM2}'s currents decrease when their respective base-emitter/source-gate terminal capacitors short, which happens at relatively high frequencies. As a result, the zero C_Z introduces appears because, while C_Z decreases the impedance at v_{EA} past p_X, the series resistance that remains when C_Z is a short circuit (i.e., R_Z in series with Q_C's emitter or diode-connected M_{PM1}'s resistance) is frequency independent at $R_Z + 1/g_{mNC}$ or $R_Z + 1/g_{mPM}$, where

$$\left.\frac{1}{C_Z}\right|_{p_X \approx \frac{1}{2\pi R_{EA} C_Z}} = R_{EA} + R_Z + \frac{1}{g_m} \approx R_{EA} \tag{7.56}$$

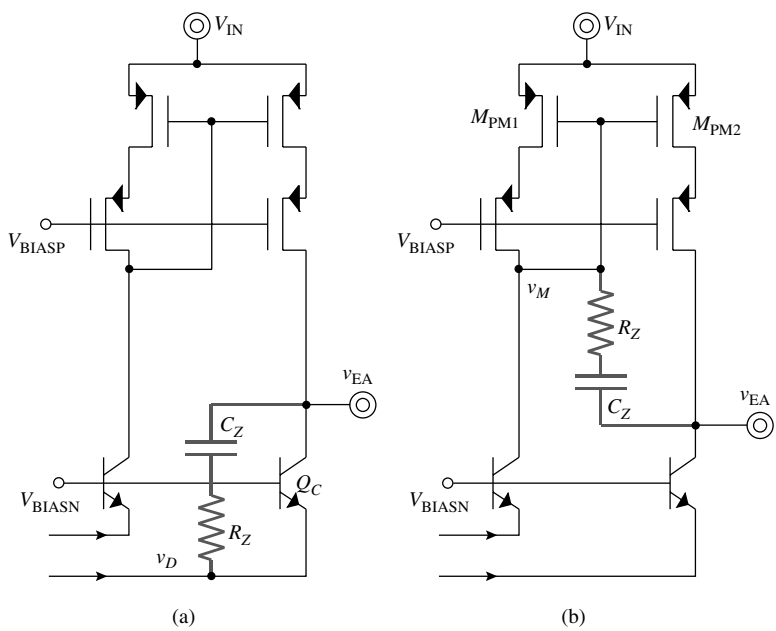


Figure 7.21 Integrating pole-zero pairs into the error amplifier.

and

$$\left.\frac{1}{C_Z}\right|_{z_X \approx \frac{1}{2\pi\left(R_Z+\frac{1}{g_m}\right)C_Z}} = R_Z + \frac{1}{g_m} \tag{7.57}$$

which means $R_Z + 1/g_m$ in z_X offsets p_X's effect on the phase response of the circuit. Note the purpose of R_Z is to increase design flexibility in $1/g_m$ and better adjust z_X's placement.

7.5.4 System

Load Regulation

An alternative to increasing loop gain (which would otherwise compromise phase margin) for the sake of dc accuracy is to shift steady-state output V_O with just enough voltage to cancel the effects of steady-state changes in output load current I_O. Since load regulation is a manifestation of the regulator's effective closed-loop output resistance $R_{O.CL}$, which in turn depends on loop gain (and systematic input-referred offset performance), I_O reduces V_O by roughly $I_O R_{O.CL}$ so the goal would be to increase V_O by the same amount I_O reduces it. One way of achieving this shift, while taking advantage of current-sensing transistor M_{PS} that buffer A_{BUF} (from Chap. 6), bulk-boosted power transistor M_{PO} (from earlier in this chapter), and the overcurrent protection (OCP) circuit (from later in Chap. 8) share, is to insert a series 50–300 Ω resistor R_{OS} between the regulator's ground terminal (now V_{OS}) and ground, as shown in Fig. 7.22. Because quiescent current I_Q is low during light loading conditions (at maybe less than 5–10 μA) and M_{PS} raises I_Q as I_O increases, R_{OS}'s ohmic drop and, in consequence, a fraction of V_O also increase with I_O:

$$\begin{aligned} V_O &= V_{REF} - I_O R_{O.CL} + V_{OS} = V_{REF} - I_O R_{O.CL} + I_Q R_{OS} \\ &\approx V_{REF} - I_O R_{O.CL} + \left.\frac{I_O R_{OS}}{A}\right|_{R_{OS}=AR_{O.CL}} = V_{REF} \end{aligned} \tag{7.58}$$

Equating R_{OS} to $AR_{O.CL}$, where A is the mirror-gain ratio between M_{PS} and M_{PO}, cancels the load-regulation effects of the feedback loop, reducing V_O to V_{REF} without any I_O effects. Note sensor Q_{P1}-Q_{P2}, as in Chap. 6, ensures the mirroring accuracy between M_{PS} and M_{PO} is acceptable during dropout conditions (by equating their respective drain-source voltages) and C_{OS} shunts displacement current to ground to mitigate the impact transient variations in I_Q produce on v_O through R_{OS}.

The key advantage of the circuit presented in Fig. 7.22, outside of improving dc accuracy, is low impact. First, R_{OS} requires a negligible

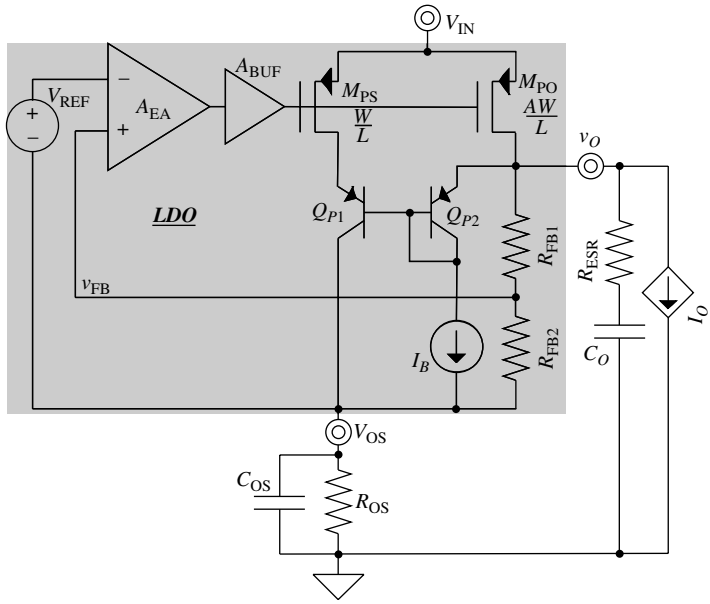


Figure 7.22 Offsetting gain-dependent load-regulation drop $I_O R_{O.CL}$ in output v_o with load-dependent shift V_{OS} or $I_O R_{OS}/A$ through what amounts to a floating ground terminal.

amount of silicon real estate. Second, which is arguably more important, R_{OS} has little to no impact on power efficiency because (1) I_Q already increases with I_O and (2) efficiency, in spite of this increase, remains high throughout the entire load range. Third, R_{OS} has no noticeable effects on frequency response and the phase margin that results because the ac signals R_{OS} introduces are common mode with respect to the entire circuit; that is, as V_{REF} shifts, so do all signals in A_{EA} and A_{BUF}. Figure 7.23*a* and *b* demonstrates how the load-enhancing resistor (i.e., R_{OS}), while having inconsequential effects on frequency response, decreases load-dependent variations in V_O from 55 mV to less than 5 mV.

From a practical standpoint, matching and tracking R_{OS} to $AR_{O.CL}$ is difficult. The least costly approach with respect to power and test time is to "match" them nominally, that is to say, to size R_{OS} to cancel the systematic effects of $R_{O.CL}$ and accept the tolerance and random variations that result, reducing, for example, a 50–60-mV drop down to ±10–15 mV. Increasing test time and silicon area to include two or three bits of trim for R_{OS} is another alternative, reducing the load-regulation variation to maybe less than ±5 mV.

The biggest disadvantage of this entire method is it requires a dedicated reference V_{REF}, since shifting V_{REF} with V_{OS} does not serve the needs of other analog circuits in the system that require an

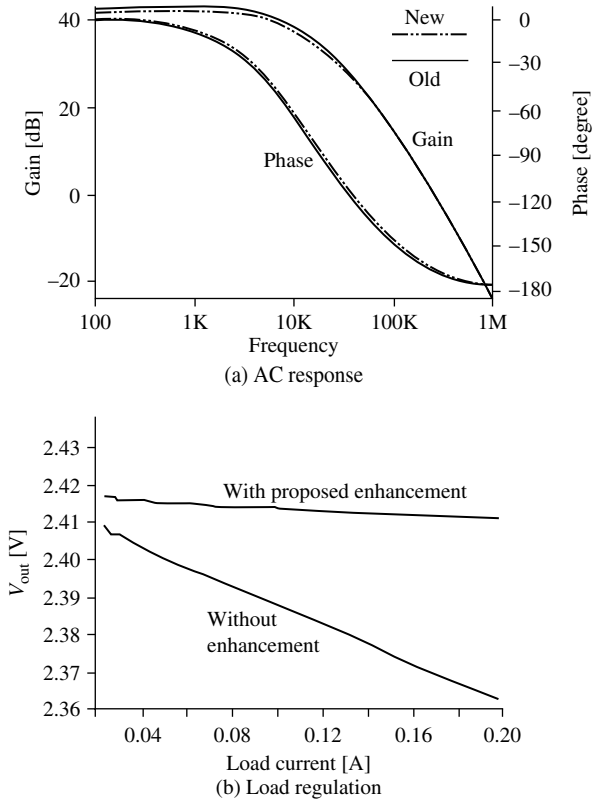


FIGURE 7.23 (*a*) Frequency and (*b*) load-regulation measurements of an LDO regulator with and without load-enhancing resistor R_{OS}.

accurate reference. Zero-load quiescent current I_{Q0} also produces a systematic offset across R_{OS} and in V_O that may be on the order of 0.5–2 mV so not only should V_{REF}'s trimming procedure account for this variation but I_{Q0} should also be low to begin with. Floating the ground of the reference-setting components alone, as shown in Fig. 7.24, effectively decreases the fraction of I_{Q0} that flows through R_{OS} and the systematic zero-load offset I_{Q0} generates. The only shortcoming here is variations in v_{OS} are no longer common mode to the circuit and A_{EA}, A_{BUF}, M_{PS}, Q_{P1}, R_{OS}-C_{OS}, and V_{REF} now constitute a positive-feedback path. Fortunately, the positive-feedback gain is considerably lower than that of its negative-feedback counterpart so its impact is normally manageable.

Transient Response

In light of all the stability complexities and issues surrounding the linear regulator, looking outside the loop to improve transient

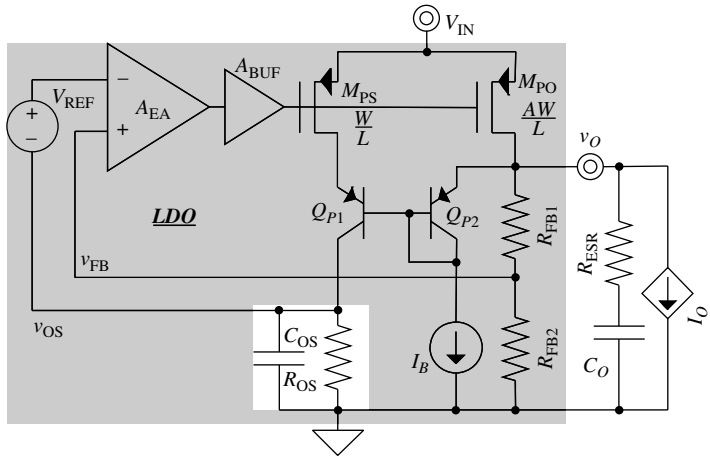


Figure 7.24 Integrating load compensating offset voltage v_{OS} into the regulator while decreasing the parasitic impact its zero-load quiescent-current flow has on accuracy.

response offers inherent value. Appending a local high-bandwidth shunt-feedback loop to output v_O, for instance, as generally illustrated in Fig. 7.25*a*, improves transient response because the faster local-feedback network responds quicker to changes in load than the more complicated regulator loop. To ensure the additional loop has negligible effects on dc accuracy, however, its low-frequency gain must be kept well below that of the regulator so that the main loop continues to enjoy control over v_O at low to moderate frequencies. The additional loop should also demand little to no current, if possible, to mitigate its negative impact on the operational life battery-powered systems achieve.

Generally, under similar current-density constraints, more devices in a feedback loop necessarily decrease the overall bandwidth of the circuit because each device introduces additional energy-shunting nodes to the ac signal path, which is why compactness and speed go hand in hand. As a result, the fastest possible loop, from an architectural point of view, is the one already present in a transistor, as in a common-drain/collector or common-gate/base configuration. As it applies to the local-feedback loop in question, Fig. 7.25*b* shows illustrative embodiments of how to bias and append these single-transistor loops to v_O. Source current i_+, for example, automatically increases when v_O decreases in response to positive load dumps, when load current suddenly rises, because Q_{N+}'s base-emitter voltage $v_{BE.N+}$ increases immediately with reductions in v_O. This is possible because, while bias current I_{NB} and diode-connected Q_{NB} establish Q_{N+}'s bias base voltage V_{N+} with respect to v_O, R_{NB}, and C_{NB} suppress ac variations emanating

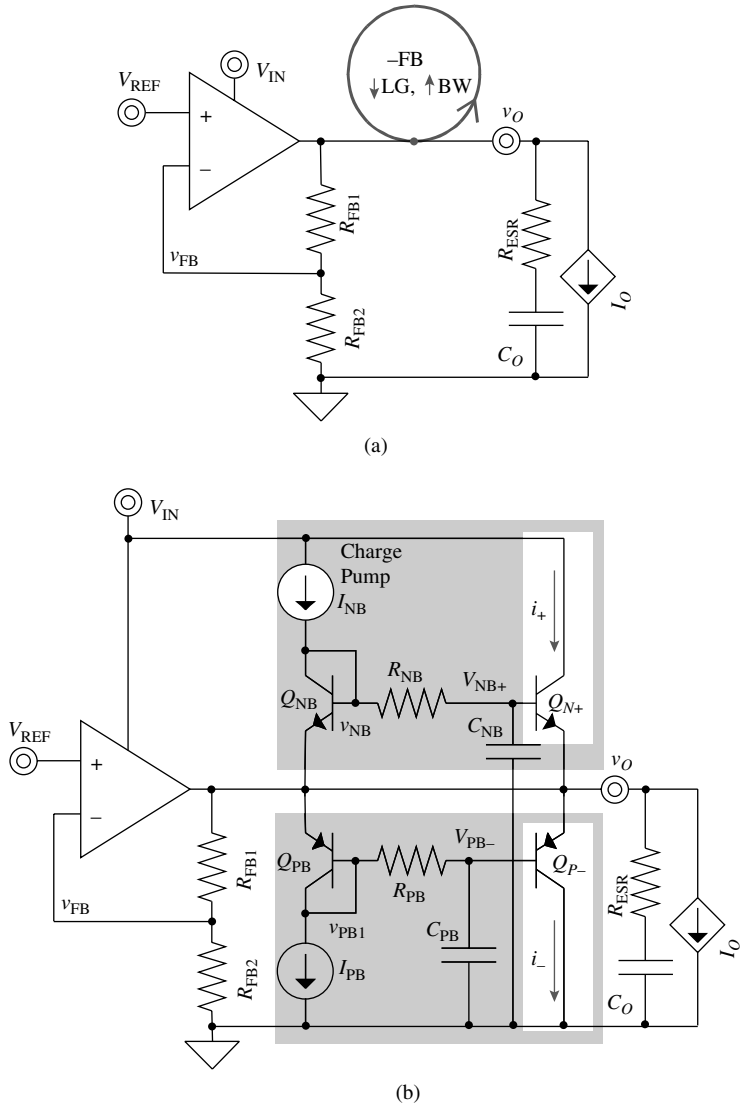


FIGURE 7.25 (*a*) Local shunt-feedback loops and (*b*) single-transistor embodiments aimed at improving the transient-response performance of an LDO regulator.

from v_{NB}. In other words, C_{NB}, for all practical purposes, fixes V_{N+} against ground so transient changes in v_O and coupled variations in v_{NB} have little effects on V_{N+} and maximum impact on $v_{BE.N+}$.

As is often the case in design, the circuit presented in Fig. 7.25*b* also poses some challenges. Increasing Q_{N+}'s transient efficacy, for

example, amounts to increasing its emitter area (i.e., increasing its output current capabilities), which means either the mirror gain between Q_{NB} and Q_{N+} increases and Q_{N+} sources a larger dc biasing current into v_O (i.e., lower efficiency) or the silicon area they require doubles with respect to the larger Q_{N+} (i.e., higher cost). What is more, the voltage across I_{NB} and Q_{NB}, when using a conventional current source, is roughly 0.7–1.1 V across process and temperature, which negates the low-dropout features of an LDO regulator; in practice, I_{NB} is a charge pump.

Transistor Q_{P-} and its associated biasing network complement Q_{N+}'s circuitry, sinking additional transient current i_- from v_O during negative load-dump conditions: in response to quick reductions in load current. However, in applications demanding output voltages exceeding 1.2 V (roughly), I_{PB} need not be a charge pump because v_O is sufficiently high to accommodate the voltage drop across the output of a conventional current mirror. Filter capacitors C_{NB} and C_{PB} are both attached to ground because connecting either of them to v_{IN} would otherwise inject supply ripple into v_O through the voltage-following configurations Q_{N+} and Q_{P-} comprise.

Power-Supply Rejection

When compared to switching power supplies, linear regulators are lossy because they conduct the full dc load across the power pass device (i.e., M_{PO}), which means the ohmic conduction loss dissipated in M_{PO} is at least $(V_{IN} - V_O)I_{LOAD}$. They nevertheless continue to be essential in, for example, among other applications, battery-powered environments, where power losses are critical, because (1) they do not inject switching noise and (2) they suppress noise already present in the supply. As a result, designers often use a switching regulator to strategically drop or boost the input supply "efficiently" and subsequently cascade an LDO regulator to "reject the noise" the switcher injects.

The challenge in using an on-chip LDO regulator for this purpose is the absence of large off-chip capacitors (which shunt noise and help compensate the negative-feedback loop) because smaller capacitors increase the sensitivity of the circuit to noise in the 0.1–10 MHz range, where dc-dc converters normally switch. The fundamental problem is the dominant low-frequency pole in internally compensated regulators normally resides inside the circuit, away from the output, so the loop tends to exhaust its transconductance gain near 100 kHz (as discussed in Chap. 5), beyond which point power-supply rejection (PSR) is most critical. A series RC low-pass filter, as shown in Fig. 7.26*a*, with a corner frequency of roughly 10–100 kHz helps attenuate noise in the region of interest, except filter resistor R_F is in the power path and consequently incurs considerable ohmic losses. Increasing the filter capacitance, incidentally, to compensate for acceptably lower resistance values (with respect to power losses) is

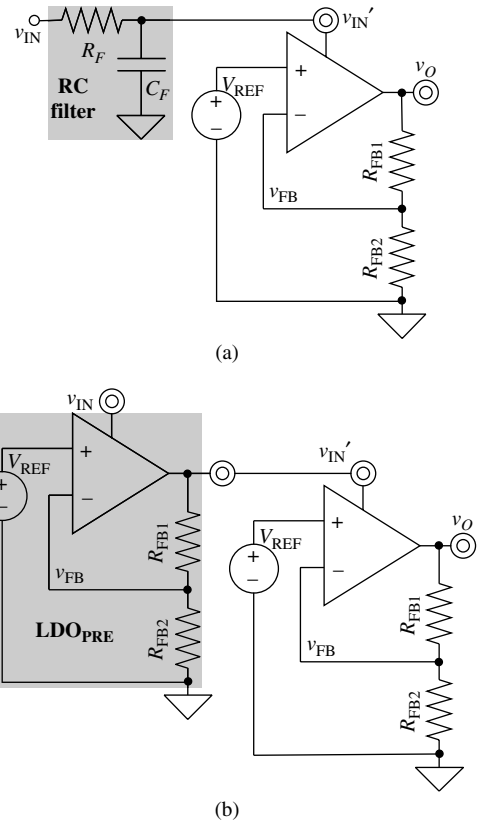


FIGURE 7.26
Increasing power-supply rejection (PSR) by passively and actively filtering the input supply with series (*a*) RC networks and (*b*) other LDO regulators.

not feasible with small on-chip capacitors. Alternatively, preregulating the supply with another regulator, as LDO_{PRE} exemplifies in Fig. 7.26*b*, further suppresses supply noise, but again, only up to 100 kHz, not where rejection is most needed. What is more, as in the RC case, LDO_{PRE}'s power device M_{PO} is also in the power path so dropout voltage and ohmic losses degrade.

Another way of improving PSR, from a conceptual perspective, is by increasing the series impedance between input supply v_{IN} and output v_O, since a voltage-divider network between v_{IN} and v_O ultimately prescribes how supply gain A_{IN} behaves across frequency (as discussed in Chap. 5). Inserting a series cascode transistor, for example, as M_{NC} portrays in Fig. 7.27, attenuates A_{IN} and increases PSR, except M_{NC} is in the power path so dropout voltage and ohmic losses also increase. Neglecting M_{NC}'s power loss, for the moment, M_{NC} also injects whatever noise is present at its gate to the LDO circuit, which is why a series low-pass RC network filters the gate. Note that setting 10–100 kHz RC corner frequencies is feasible in this setup because

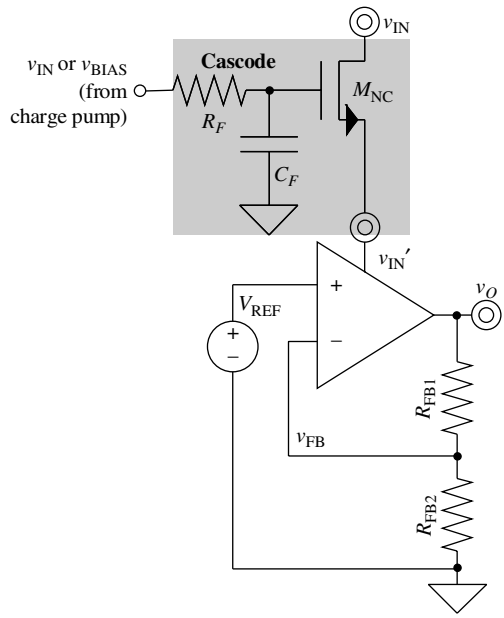
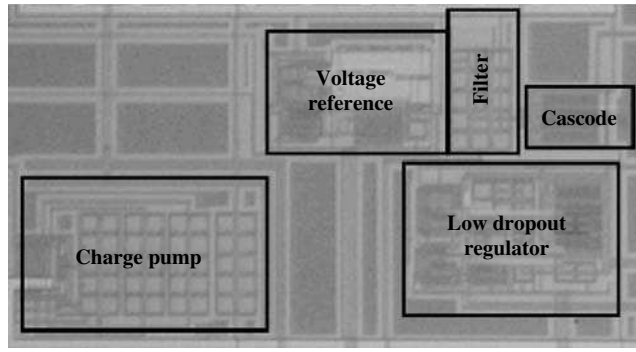


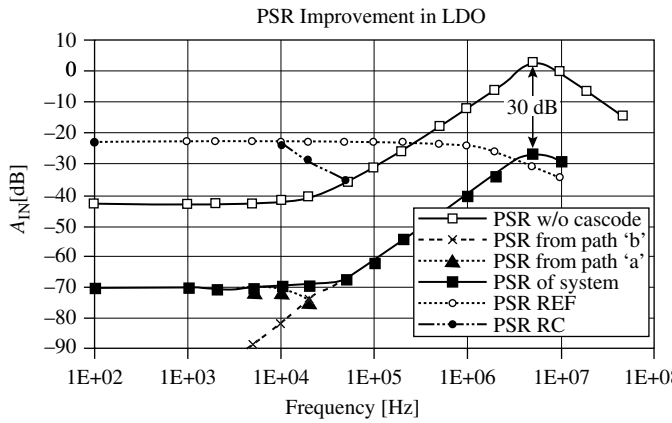
Figure 7.27
Improving PSR by increasing the impedance from the output to the input supply with a cascode transistor.

resistor R_F no longer has to subscribe to stringent size restrictions, now that R_F is not in the power path, so a 250-kΩ–25-pF filter, just to cite an example, which produces a 26-kHz corner frequency, is viable and practical.

Going back to power losses, M_{NC} should not sustain drain-source voltages below its saturation level of $V_{DS.NC(sat)}$ because M_{NC}'s drain-source resistance would otherwise decrease to the point M_{NC} no longer offers any advantages with respect to PSR. Even if the power loss associated with $V_{DS.NC(sat)}$ were acceptable, biasing the device just above saturation requires a gate-bias voltage that exceeds the supply, which is why a practical implementation of this circuit uses a charge pump to bias M_{NC}. Note the switching frequency of the charge pump should remain well above the filter's corner frequency to ensure none of its switching noise and related harmonics reach M_{NC}'s gate. M_{NC}'s gate voltage should also be high enough to account for bulk effects, if any, and the large load current it conducts. For reference, Fig. 7.28 illustrates the chip photograph and supply-gain measurements of a 0.5-μm CMOS LDO regulator with and without a cascode and its corresponding 3 kHz low-pass RC filter biased from a 10-MHz charge pump with 100 mV_{PP} of ripple. The results show the charge-pumped cascode decreases A_{IN} (and increases PSR) by approximately 30 dB across the entire frequency band tested.



(a)



(b)

FIGURE 7.28 (*a*) Chip photograph and (*b*) supply-gain measurements of a 0.5-μm CMOS LDO regulator with and without a cascode biased through a 3-kHz RC filter from a 10-MHz charge pump with 100 mV_{PP} of ripple.

7.6 Summary

The underlying aim of this chapter was to illustrate how to assemble the integrated circuits developed in Chap. 6 into a working system that addresses the overall design objectives described in Chap. 1. To this end, developing and applying a robust and efficient compensation strategy to ensure the circuit is stable across all operating conditions is one of the first and most important steps in this part of the design process. The basic strategy is to ensure only one dominant low-frequency pole exists, only one secondary pole resides near unity-gain frequency f_{0dB}, and all other parasitic poles and right-half-plane zeros land a decade past f_{0dB}. Left-half-plane zeros then

help extend f_{0dB} slightly by mitigating the avalanching effects several poles have on phase margin. Ultimately, however, even after optimization, low-power constraints (through their impact on $1/g_m$ resistances) limit the overall bandwidth of the system (i.e., f_{0dB}), which means off- or on-chip capacitors must source and sink load dumps whose rising and falling rates exceed the speed of the regulator.

The reference is another consideration, and although regulators often rely on external reference circuits, incorporating the workings of V_{REF} into the feedback loop is not uncommon, especially in standalone applications and segmented, power-aware environments. The first step in this process is to integrate the proportional-to-absolute-temperature (PTAT) current generator into the feedback loop, ensuring the positive-feedback path it presents has lower gain than the negative counterpart does. Additionally, a start-up circuit must guarantee the reference reaches its desired state, even when initially latched otherwise. Achieving temperature independence then amounts to introducing a complementary-to-absolute-temperature (CTAT) component—this typically appears in the form of a series diode voltage. Decreasing the temperature dependence beyond this point is often not worth the trouble because the random piezoelectric effects of the fillers in the plastic package on the circuit overwhelm most second-order variations in the reference. Nevertheless, decreasing the temperature coefficient below first-order levels is best achieved by exploiting the second-order components inherent to the circuit, such as differences in supposedly matched collector-emitter or drain-source voltages.

The importance of dropout, power, speed, accuracy, and power-supply rejection (PSR) in linear regulators cannot be understated, which is why developing performance-enhancing circuit strategies enjoys so much attention in industry and research circles. However, while improving and complicating a circuit tend to go hand in hand, increasing circuit complexity often slows the circuit, demands more power, requires more silicon area, and increases risk. The trick in designing robust and elegant solutions is to find efficient means of exploiting and leveraging the transistors and devices that already exist in the circuit, limiting the extent risk and tradeoffs plague the design. Driving the bulk of a power MOS switch, for instance, may increase its effective gate drive but not without increasing the propensity for parasitic BJTs to conduct current. Any attempt at increasing PSR, just to cite another example, also tends to increase quiescent power in the form of dropout. Nevertheless, finding a sensible and optimal balance between conflicting specifications, design objectives, and complexity *is* possible, especially when understanding the application and how the regulator affects it.

In general, designing linear regulator ICs is the art of using semiconductor technologies (from Chap. 2) to build circuits (like those in Chaps. 3 and 6) that regulate and condition energy and power (as prescribed in Chaps. 4 and 5) according to the system demands driving

state-of-the-art applications impose (as described in Chaps. 1 and 7). Understanding the application and the constituent technologies that drive the design is extremely important in this process, especially when attempting to extend battery life and improve regulating performance. To this end, this chapter in many ways concludes the discussion on the design of linear regulator ICs, since the textbook has, at this point, presented almost all the expertise necessary to design one. The only aspects left to discuss are circuit protection (as in thermal, overcurrent, etc.) and testing, which the next chapter briefly addresses.

CHAPTER 8
IC Protection and Characterization

Although understanding the performance objectives of a linear regulator and designing an integrated circuit (IC) that aims to fulfill those goals (Chaps. 1–7) are intrinsic and involved, neglecting to protect the regulator against damaging (and often catastrophic), unintended (but nonetheless realistic) events renders the IC ineffectual and the system it supports inoperable and susceptible to more injurious conditions. In preventing these ill-fated circumstances, the first line of defense is to protect the IC against such undesired, yet realistic conditions at the circuit level, considering the power levels involved are relatively high and consequently prone to surpass the operating and breakdown limits of the regulator. The second step in this process, after fabrication, is to comprehend the operating and performance limits of the IC and avoid overloading it with systems whose operating and power demands exceed its specified capabilities. Ascertaining these constrains experimentally is important because the three-sigma (i.e., 3σ) semiconductor models used to predict the performance of the IC are imperfect (relative to real-life die-to-die, lot-to-lot, and performance-over-time variations) and incomplete (because they lack many of the parasitic devices present in the IC). To make matters worse, simulations often exclude systematic and random parasitic effects such as substrate-noise coupling, random mismatches in supposedly matched transistors, hot spots and thermal gradients across the die, and others. So, in harmony with the design approach thus far presented in the textbook and the product- and prototype-development cycle it seeks to support, this chapter discusses circuit protection (as considered and addressed before fabrication) and characterization (once the IC is fabricated).

8.1 Circuit Protection

In basic terms, the purpose of on-chip circuit protection is to avoid exposing any part of the IC to conditions that surpass its breakdown limits. Bearing in mind power pass transistor S_O conducts all the load

current, S_O is most likely to exceed its *safe-operating area* (SOA), which means the IC must prevent S_O from reaching maximum rated limit $P_{SO(max)}$. It is no surprise, then, that *overcurrent protection* and *thermal shutdown* features are popular in regulator ICs. Similarly, *reverse-battery* and *electrostatic-discharge* (ESD) protection are also important and therefore included in most, if not all, of today's systems.

Protection, however, should not risk or interfere with the functionality or performance of the regulator so it should not only prove reliable but also remain off and "transparent" during normal operating conditions, and engage only when needed, on demand. Transparency, among other things, also implies drawing little to no quiescent current from the supply, which is why trip-point accuracy for these functions is normally poor. Fortuitously, because regulation and power performance sells more than protection (especially in the battery-powered segment), regulator ICs are often overdesigned (with respect to $P_{SO(max)}$) for the sake of output accuracy and efficiency, which means margin exists to accommodate for variations in the protection trip points. Margin is also important to avoid inadvertent transient glitches in the system, and deglitch filter functions help in that regard.

8.1.1 Overcurrent Protection

Fixed Overcurrent Protection

Overcurrent protection (OCP), whose aim is to protect S_O, is another term for *short-circuit* and *overload* circuit protection. Its objective is to keep the power dissipated across S_O below its maximum-rated limit of $P_{SO(max)}$:

$$P_{SO} \equiv V_{SO} I_{SO} = (V_{IN} - V_{OUT}) I_{OUT} \le P_{SO(max)} \equiv V_{IN} I_{SO(max)} \tag{8.1}$$

where V_{IN} is the maximum voltage applied to S_O during short-circuit conditions, when V_{OUT} is 0 V and S_O is exposed to the entire supply. Note $P_{SO(max)}$ also accounts for worst-case temperature conditions so it includes the effects of high ambient temperature T_A (e.g., 125°C) and the resulting (and even higher) junction temperature T_J (e.g., 130°C)—how much higher T_J is over T_A depends on the power dissipated across S_O and the package's thermal impedance θ_P. Considering the accuracy of OCP limit I_{OCP} is relatively poor (because quiescent current must remain low for high power efficiency), the nominal value of its worst-case implied power $P_{OCP(max)}$ falls below $P_{SO(max)}$,

$$P_{OCP(max)} \equiv V_{SO(max)} I_{OCP} \le V_{SO(max)} I_{SO(max)} \equiv P_{SO(max)} \tag{8.2}$$

as shown in Fig. 8.1*a*, where $V_{O.TAR}$ refers to the regulator's target output voltage and the regulator limits I_{OUT} to I_{OCP} during overload and short-circuit conditions. Similarly, maximum allowable output current $I_{OUT(max)}$ must remain below I_{OCP} to accommodate for I_{OCP} variations,

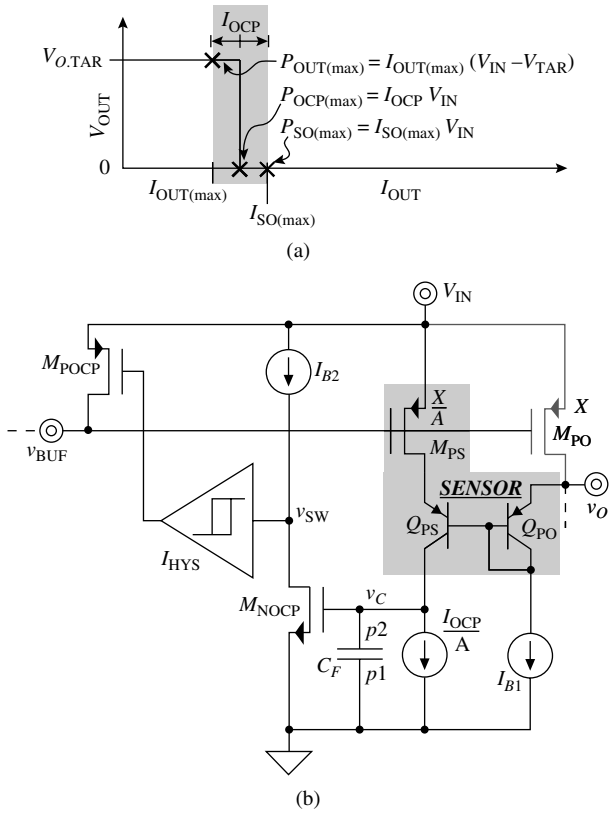


FIGURE 8.1 Fixed overcurrent protection (*a*) V_{OUT}-I_{OUT} transfer function and (*b*) sample PMOS-switch circuit embodiment.

which ultimately means the maximum specified power S_O can deliver (i.e., $P_{OUT(max)}$) is considerably below its maximum rating of $P_{SO(max)}$:

$$\frac{P_{OUT(max)}}{P_{SO(max)}} = \frac{I_{OUT(max)}\left(V_{IN} - V_{O.TAR}\right)}{I_{SO(max)} V_{IN}}$$

$$= \frac{I_{OUT(max)}}{I_{SO(max)}}\left(1 - \frac{V_{O.TAR}}{V_{IN}}\right) < \frac{I_{OCP}}{I_{SO(max)}}\left(1 - \frac{V_{O.TAR}}{V_{IN}}\right) \qquad (8.3)$$

Figure 8.1*b* illustrates a sample circuit embodiment of the fixed OCP graph shown in Fig. 8.1*a*. Current sensor PMOS transistor M_{PS} senses and sources a linear fraction of output PMOS device M_{PO}'s current I_{OUT} (i.e., I_{MPS} is I_{OUT}/A). Sensor BJTs Q_{PO} and Q_{PS}, as in the buffers discussed in Chap. 6, impress M_{PO}'s drain-source voltage across M_{PS} to ensure both devices remain in similar regions of operation (for higher mirror accuracy). As a result, when I_{MPS} approaches I_{OCP}/A, I_{MPS} raises

control voltage v_C and M_{NOCP} starts pulling switching node v_{SW} down from V_{IN}. When I_{MPS} nears I_{OCP}/A, digital hysteretic inverter I_{HYS} switches and causes M_{POCP} to pull buffer node v_{BUF} closer to V_{IN}, shutting off M_{PO} in the process. In the case I_{LOAD} remains at I_{OCP}, M_{PO} may continue to engage and disengage about the I_{OCP} threshold, if short-circuit conditions persist, with a frequency that is dependent on the time constant associated with the OCP loop and on the hysteresis in the inverter, which is why adding hysteresis to I_{OCP} is important, to avoid oscillations. Filter capacitor C_F, besides decreasing this frequency, prevents extraneous noise in v_C from inadvertently engaging M_{NOCP} and causing false alarms and glitches in the system.

Fold-Back Overcurrent Protection

In the portable space, over-designing the linear regulator to source less power than it can actually handle (i.e., $P_{O(max)}$ is considerably less than $P_{SO(max)}$) is often a natural by-product of the dropout specification, which forces the size of the power device to be larger than its power rating demands. Many higher power applications, however, do not enjoy this providential circumstance, which is where fold-back current limiting finds a niche. The problem with a fixed OCP limit is the V_{SO} voltage that sets $P_{O(max)}$ when operating normally (i.e., V_{SO} is $V_{IN} - V_{O.TAR}$, which can be 300 mV) can be vastly smaller than the one defining $P_{OCP(max)}$ during short-circuit conditions (when V_{SO} is V_{IN}, which can be 6 V). One way to extend $P_{O(max)}$, as shown in Fig. 8.2*a*, is to allow I_{OUT} to push S_O closer to $P_{SO(max)}$ during normal operation and, once overloaded past prescribed initial OCP limit $I_{OCP(ini)}$, decrease (i.e., *fold back*) I_{OCP} with reductions in V_{SO}, as V_{OUT} decreases and the IC reaches short-circuit conditions. In other words, the driving objective in folding back the OCP limit is to keep S_O's power P_{SO} approximately constant across the entire overload range. In practice, $P_{OUT(max)}$ should precede $P_{OCP(max)}$ because of possible variations in I_{OCP} (Fig. 8.2*a*), as before; but even then, $P_{OUT(max)}$ is now considerably higher and closer to $P_{SO(max)}$ when compared against its fixed counterpart, as $I_{OUT(max)}$ also is with respect to $I_{SO(max)}$:

$$\frac{P_{OUT(max)}}{P_{SO(max)}} = \frac{P_{OCP(max)} - \Delta I_{OCP} V_{SO(max)}}{P_{SO(max)}} < \frac{I_{OCP} V_{SO(max)}}{I_{SO(max)} V_{SO(max)}}$$

$$= \frac{I_{OCP}}{I_{SO(max)}} < 1 \tag{8.4}$$

where $\Delta I_{OCP} V_{SO(max)}$, or equivalently, ΔP_{VAR} is the power margin included to accommodate variations in I_{OCP}.

To implement a fold-back limiter, the OCP circuit must sense both i_{OUT} and v_{OUT} to assert OCP when i_{OUT} first reaches initial limit $I_{OCP(ini)}$ and decrease I_{OCP} with reductions in v_{OUT}, as the circuit approaches short-circuit conditions. Unfortunately, relative to fixed current

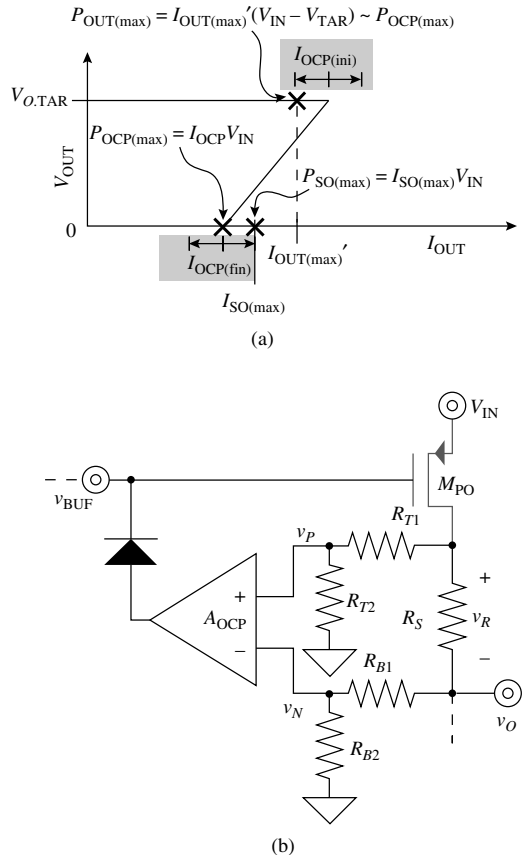


FIGURE 8.2
Fold-back overcurrent protection (*a*) V_{OUT}-I_{OUT} transfer function and (*b*) sample circuit embodiment.

limiting, sensing v_{OUT} represents an additional complexity that may add quiescent current, dropout voltage, risk, silicon area, and/or cost. Nevertheless, one way to sense both i_{OUT} and v_{OUT} simultaneously is to monitor the voltage across a series resistor R_S in the output path of the regulator with respect to ground (via a pair of resistive voltage dividers), as illustrated in Fig. 8.2*b*. As shown, OCP engages when amplifier A_{OCP}'s input differential voltage $v_p - v_N$ exceeds 0 V; otherwise, A_{OCP} attempts (in vain) to pull current from blocking diode D_{OCP}.

By setting top resistor divider R_{T1}-R_{T2}'s gain K_T to be lower than R_{B1}-R_{B2}'s gain K_B (i.e., K_T is less than K_B) and using resistor values that substantially exceed R_S, v_N (which is initially higher than v_p at light loads) decreases faster (i.e., with a higher gain) than v_p in response to a rising i_{OUT} (i.e., increasing v_R); that is,

$$\begin{aligned} v_P - v_N &= (v_{OUT} + I_{OUT}R_S)K_T - v_{OUT}K_B \\ &= v_{OUT}(K_T - K_B) + I_{OUT}R_SK_T \end{aligned} \tag{8.5}$$

and

$$I_{OCP} = I_{OUT}\Big|_{v_P = v_N} = v_{OUT}\left(\frac{K_B - K_T}{R_S K_T}\right) \tag{8.6}$$

As a result, when v_N decreases past v_P (i.e., $v_P - v_N$ exceeds 0 V), OCP engages and the loop A_{OCP} creates regulates I_{OUT} to I_{OCP}, and as v_{OUT} decreases, so does I_{OCP}. Note the loop is analog in nature and must therefore remain stable. In practice, A_{OCP} may share some of its transistors with regulator buffer A_{BUF} and/or error amplifier A_{EA}. Note, as before, including some hysteresis in the OCP trip point mitigates the circuit's propensity to oscillations when I_{OUT} is just at OCP limit (i.e., when I_{OUT} equals I_{OCP}).

8.1.2 Thermal Shutdown

Junction temperature T_J, as it turns out, is sometimes an indirect measure of i_{OUT} because, barring the existence of other supplies on the chip, T_J increases with the power dissipated in the regulator (i.e., P_{REG}), almost all of which is concentrated in P_{SO}:

$$T_J = T_A + P_{REG}\theta_P = T_A + P_{SO}\theta_P = T_A + I_{OUT}V_{SO}\theta_P \tag{8.7}$$

where as before, θ_p is the thermal impedance the package. As a result, when considering the regulator on its own, thermal shutdown and the OCP function protect the IC from the same stressing conditions, which is why implementing both functions on a stand-alone linear regulator can be redundant. With multiple regulators and power devices on the same die, however, the situation changes because hot spots and thermal time constants de-correlate T_J from a regulator's i_{OUT}.

Because S_O is often overdesigned (with respect to power) to meet low-dropout specifications, the purpose of thermal protection is not always to protect S_O but to keep the package from melting, as is typically the case in portable applications. Therefore, since the melting point of plastic packages T_P is near 170°C, thermal trip point T_{SHUT} must fall below T_p by a margin, to maybe 150°C, considering, as with I_{OCP}, low-cost requirements such as low quiescent current, limited silicon real estate, short test time, and others induce inaccuracies and variations in T_{SHUT}. Similar to the OCP case, thermal shutdown should include hysteresis around T_{SHUT} to prevent the circuit from oscillating uncontrollably about T_{SHUT}. In practical terms, hysteresis window T_{HYS} around T_{SHUT} means the circuit trips at T_{SHUT} but recovers only when T_J falls below T_{SHUT} by T_{HYS} (i.e., T_J is $T_{SHUT} - T_{HYS}$).

Perhaps the most practical and predictable means of establishing a temperature trip point is to sense when a proportional-to-absolute temperature (PTAT) voltage V_{PTAT}, which is well-modeled and stable, decreases past its increasing complementary-to-absolute temperature (CTAT) counterpart V_{CTAT}, which is simply a base-emitter voltage V_{BE} or

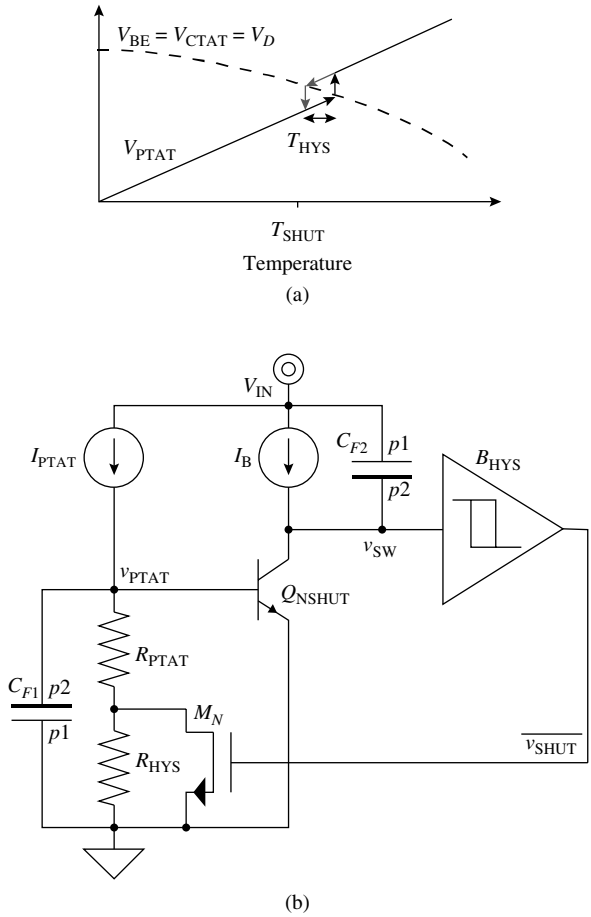


FIGURE 8.3 Thermal shutdown (*a*) trip-point graph and (*b*) sample circuit embodiment.

a diode voltage V_D, as depicted and implemented in Fig. 8.3. With respect to the sample circuit embodiment shown in Fig. 8.3*b*, while operating within the specified temperature range, v_{SHUT} is low and its complement is high so M_N shorts hysteresis resistor R_{HYS} and sensing PTAT voltage v_{PTAT} is only the voltage across R_{PTAT}. Once v_{PTAT} is sufficiently high to induce Q_{NSHUT} to sink bias current I_B and pull v_{SW} past the digital buffer's threshold, shutdown asserts (i.e., v_{SW} rises and v_{SHUT} falls) and M_N shuts off. Allowing I_{PTAT} to flow through R_{HYS} increases v_{PTAT} and pushes the circuit further into its shutdown regime so, when temperature again decreases, R_{SHUT}'s voltage must decrease below its assertion level to disengage the circuit, establishing the desired T_{HYS}. Filter capacitors C_{F1} and C_{F2}, as in OCP, prevent (i.e., deglitch) extraneous noise signals injected into v_{PTAT} and v_{SW} from inadvertently asserting shutdown.

8.1.3 Reverse-Battery Protection

The purpose of reverse-battery protection, as the name implies, is to shield the regulator from the effects of sustained reverse-battery conditions. The most efficient method of achieving this objective (with respect to quiescent current, silicon area, and electrical interference in the form of noise and undesired voltage drops) is to altogether avoid the condition by, for example, "keying" the battery holder with plastic guards and/or guides so the battery can only fit when properly oriented. This kind of protection is possible and popular in many lithium-ion (Li-Ion) applications such as cellular phones, portable digital assistants (PDAs), palm pilots, MP3 players, and the like. Coin cells, unfortunately, are difficult to key so they often include protection in the electrical domain, which usually degrades some of the device's electrical performance.

As in OCP and thermal shutdown, electrical reverse-battery protection should remain off and "transparent" (but alert) during normal operating conditions, and assert only when needed on demand. Transparency, as before, implies quiescent current i_Q, series voltage drops in the supplies v_S, and reverse current i_R should all be zero, or at least near zero. Note that to sustain these extreme adverse conditions for any length of time, the semiconductor devices confronting those conditions must be capable of handling high power and/or high electric fields. These devices must therefore be large and heavily guard-ringed, much like power pass transistor S_O, and consist of transistors and diodes whose breakdown voltages are high, such as is the case with devices built with low doping-density junctions.

Functionally, reverse-battery protection either shunts current away from the regulator circuit or blocks it. Shunting current with a diode, for instance, as shown in Fig. 8.4*a*, protects the regulator, but does so at the expense of reverse-battery current i_R, which can quickly drain the battery. Blocking current, as a result, is better with respect to battery life, but not forward efficiency, as blocking often introduces a series voltage v_S in one of the supplies during normal operating conditions,

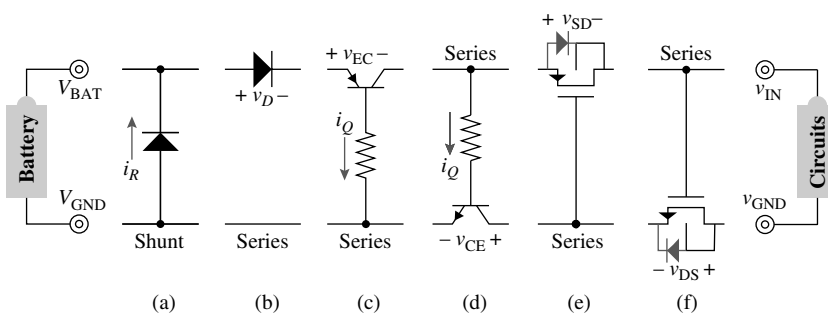


Figure 8.4 (*a*) Shunt diode, (*b*) series diode, (*c*) PNP BJT, (*d*) NPN BJT, (*e*) PMOSFET, and (*f*) NMOSFET reverse-battery protection circuits.

which dissipates power (although not as much as shunt protection during reverse-battery conditions) and decreases the effective supply to the circuit (i.e., decreases v_{IN}), adversely impacting headroom performance.

A series diode, for example, blocks reverse-battery current (as in Fig. 8.4*b*) but introduces diode voltage v_D in series with battery voltage V_{BAT}. Series transistors, as illustrated in Fig. 8.4*c* through *f*, are better blockers because their respective series voltages (i.e., v_{CE} or v_{DS}'s) are lower, especially n-type transistors, since n-type mobility μ_N is normally 2–3 times higher than its p-type counterpart. BJTs, however, dissipate quiescent base current I_Q whereas MOSFETs normally require more silicon area. (Note MOSFET bulks are tied to their respective drains to ensure the parasitic body diode that remains is off in reverse-bias conditions during reverse-battery events.) Ultimately, irrespective of the means, unless mechanical restraints are possible, avoiding parasitic reverse current i_R, series voltage v_S, and/or quiescent current i_Q is difficult.

8.1.4 Electrostatic-Discharge Protection

Electrostatic discharges (ESD) are not entirely unlike reverse-battery conditions, except the former are short lived, occur in both directions, and the voltages involved are substantially higher. Most ESD strikes result from one of three basic mechanisms. The first and most obvious source is the human touch, after generating static charge by walking across a carpet, for example, which is modeled by what is termed the *human-body* model (HBM). Similarly, although from a different source, the *machine* model (MM) mimics the charge transferred to the IC when machines handle the chips during the bundling and shipping processes. Finally, the *charged-device* model (CDM) describes the electrical conditions that result when a precharged chip, after sliding down its plastic container (i.e., tube), touches chassis ground. Note the common thread in all these conditions is the sudden release of charge, which is why precharged picofarad capacitors with initial voltages ranging from 250 V to 5 kV and various series resistances emulate these undesired but realistic events.

The performance of ESD circuits is strongly dependent on device layout and process technology, and predicting the extent to which they protect the IC is largely empirical. ESD structures, as a result, vary significantly across process, the junctions and polysilicon materials they protect, and even semiconductor companies. What is worse, all input and output pins (I/Os) are susceptible to positive and negative ESD strikes so they all need protection.

Most ESD protection circuits engage (i.e., "trigger") via dc or ac coupling schemes and clamp their respective I/Os with forward- or reverse-biased diodes or *silicon-controlled rectifiers* (SCRs). For example, a common ESD structure used in digital ICs is a pair of forward biasing diodes between pin v_{PIN} and supply V_{IN} and ground, as shown

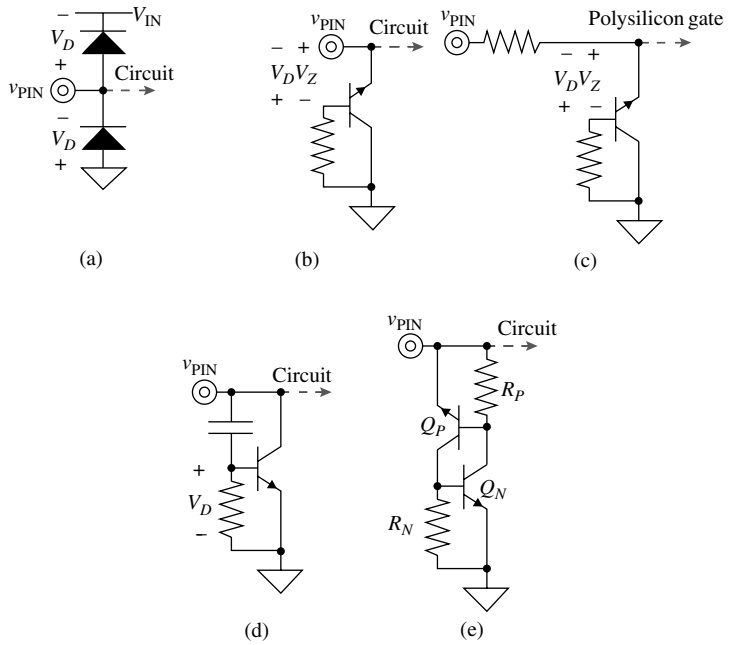


Figure 8.5 Common electrostatic-discharge protection circuits: (*a*) dc-coupled forward biasing supply diodes, (*b*) dc-coupled forward/reverse biasing Zener junction, (*c*) resistor-decoupled clamp (for polysilicon gates), (*d*) ac-coupled forward biasing BJT, and (*e*) silicon-controlled rectifier.

in Fig. 8.5*a*, where positive and negative strikes engage the supply and ground diodes, respectively, and clamp v_{PIN} to $V_{IN} + V_D$ and $-V_D$. The drawback to this approach is that the response to positive strikes relies and depends on V_{IN}'s ESD clamp, which means current circulates from v_{PIN} through the IC to V_{IN} before finally reaching ground.

Fortunately, reverse biasing a diode also shunts energy away from the pin, thereby circumventing the dreaded V_{IN} path to ground. Figure 8.5*b* shows one way to exploit this feature in a BJT, by forward biasing its base-emitter junction during negative strikes and reverse biasing them in response to positive strikes. (NPN base-emitter junctions are commonly used to build Zener diodes.) Polysilicon gates, as it turns out, are especially sensitive to CDM-like strikes so placing local Zener clamps near the gates and decoupling them from the pin via series resistors, as illustrated in Fig. 8.5*c*, typically help—since the gates do not draw dc current, series resistors do not normally present a problem during normal operating conditions. Considering ESD events are sudden and transient in nature, ac coupling the energy through a capacitor is also a viable means of engaging a transistor, as shown in Fig. 8.5*d*, where the NPN BJT steers ESD energy to ground when the strike builds sufficient voltage across its base-emitter junction to engage it. In these cases, the designer tunes (often empirically)

the resistor-capacitor time constant to match the rise and fall times expected in the strikes.

Silicon-controlled rectifiers, as illustrated in Fig. 8.5*e*, are also popular ESD structures. These circuits rely on latch up, which is a manifestation of positive feedback, to steer ESD energy away from the sensitive circuit. During normal operating conditions, for example, there is no base-emitter voltage across either BJT (in Fig. 8.5*e*) to engage the circuit. Forward biasing one of the base-emitter junctions, however, with sufficient transient energy to induce a collector current engages the complementary BJT and latches the circuit. In other words, when referring to Fig. 8.5*e*, as Q_N causes Q_P to conduct, Q_P responds by forward biasing Q_N further, accentuating the forward biasing process and causing the collector currents to increase and absorb the incident ESD energy. Bear in mind noise energy should not engage the SCR so the design should include sufficient noise margin to prevent inadvertent latch-up events from occurring.

With respect to linear regulators, specifically, noting the power transistor is often self-protecting is worthwhile. The fact is the pass device must already sustain high (load) power and channel what amounts to considerable current back to the supplies via low-impedance paths. The device is therefore guard-ringed with several n- and p-type diffusion and metallic layers and placed near the edge of the die, close to V_{IN} and the ground bus. Consider, for example, the large body diode of a power PMOSFET already protects v_{OUT} against positive ESD strikes by quickly steering energy to V_{IN}'s ESD circuit, which is normally, with respect to die placement, as alluded earlier, next to the regulator IC. If v_{OUT} is also attached to a polysilicon gate, however, as in the case of a MOS-input differential pair, a local clamp near the gate in question with a series resistor (e.g., 1–10 kΩ), as illustrated in Fig. 8.5*c*, should also be considered.

8.2 Characterization

Well-specified data sheets shield ICs from overstressing conditions as much as protection circuits do, if not more. The reality is inaccurate and/or incomplete information on the statistical performance and operating limits of the IC can only tempt end users and system designers into systematically overstressing the chip. The worst by-product of this kind of neglect is the system may not fail immediately but later, after the product using the IC reaches the market. Not only will recalling and fixing merchandise after the fact incur substantial manufacturing and personnel expenses but so will losing the reputation that attracted the customer base in the first place, as users and consumers lose confidence in the goods the company sells. Data sheets must therefore specify and include carefully tested and statistically meaningful information on the performance of the IC.

This section discusses the most common test methods used to determine the operating limits of a linear regulator. Since semiconductor profits hinge on large-scale (i.e., volume) production and sales, increasing the statistical value and confidence of the data is extremely important. Numerous ICs from several different wafers and fabrication lots are therefore fully tested across worst-case temperature and operating conditions. The underlying objective is to ascertain the regulation and power performance of the IC and its operational requirements. Although not often quoted in data sheets, the designer must also ensure the IC starts and recovers properly from worst-case power-up and power-down sequences.

8.2.1 Emulating the Load

To emulate the effects of a practical load on the regulator's loop gain and stability conditions, among other things, it is important to grasp its nature. Unfortunately, outside of the specified load-current range, the load is largely unknown and unpredictable, which means understanding its most relevant features along with all its possible manifestations is imperative. Thankfully, a Norton-equivalent model is sufficient, but only for a particular load setting, not the entire range. Consider, for instance, that while a variable current source can emulate variations in current, it does not account for its effects on parallel load resistance R_{LOAD}.

Nevertheless, given the practical limitations a laboratory environment presents, modeling the load with a simple resistor, as depicted by variable resistor R_{LOAD} in Fig. 8.6, is often easiest. The idea is to leverage the regulating performance of the circuit by allowing v_{OUT} to set its own load current i_{LOAD} with respect to R_{LOAD}; in other words, allow the ohmic voltage across R_{LOAD} (which is v_{OUT}) define i_{LOAD} (i.e., i_{OUT} is v_{OUT}/R_{LOAD}). The problem with this setup is R_{LOAD}'s small-signal resistance, while delivering a particular i_{LOAD} setting, may not be accurate. Consider, for example, the loading resistance of an op amp biased with a "supply- and output-independent" quiescent current of 10 mA (i.e., i_{LOAD}) and regulated to supply an output

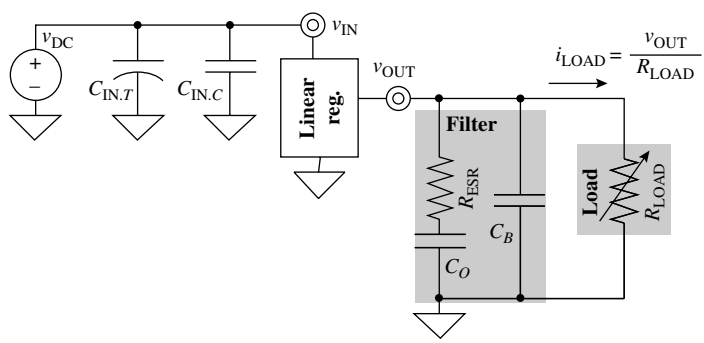


Figure 8.6 Test circuit for a linear regulator with a resistor load.

of 2 V (i.e., v_{OUT}) is not necessarily equivalent to "output dependent" R_{LOAD} (i.e., v_{OUT}/i_{LOAD}), which would otherwise indicate 2 V/10 mA or 200 Ω. In all fairness, the model of a switching digital IC, just to cite a different example, often reduces to an equivalent resistance because its current is, for the most part, proportional to its supply (i.e., i_{LOAD} is proportional to v_{OUT}), which means a purely resistive load model in this case is relatively accurate.

On the other extreme, as mentioned earlier, modeling the load with a current source does not account for variations in R_{LOAD} with respect to i_{LOAD}. It is nonetheless useful, however, to also examine this other extreme because the load is, as before, principally unknown. The circuit in Fig. 8.7, as an example, emulates what amounts to an all-active load. The op amp used (assuming it has high gain and negligible input-referred offset), irrespective of v_{OUT}, ensures the voltage across R_{SET} is v_{SOURCE} so its current, which flows through NMOS transistor M_N, is v_{SOURCE}/R_{SET}, or equivalently, user-defined i_{LOAD}. Although modeling the load in this fashion does not include the effects of a purely resistive load, implementing both setups examines the effects of R_{LOAD}'s extreme values on the circuit, especially as it pertains to loop gain and stability where R_{LOAD} may have a considerable impact.

With respect to the circuit shown, the tantalum- or electrolytic-ceramic capacitor combination $C_{IN.T}$-$C_{IN.C}$ attached to v_{IN} in both Figs. 8.6 and 8.7 ensures the supply at the regulator pin is free of noise, that is, stable and predictable. Because a ceramic capacitor introduces low ESR, $C_{IN.C}$ shunts high-frequency noise whereas $C_{IN.T}$, being that tantalum and electrolytic capacitances (and their respective ESRs) can be substantially higher, shunts low-frequency noise. Relatively high voltage sources, incidentally, supply the discrete op amp because its headroom limit is often higher than the IC regulator is. Additionally, the negative supply is substantially below ground to ensure power transistor M_N does not enter triode during heavy loading conditions, when the voltage across R_{SET} (and M_N's source voltage) is highest.

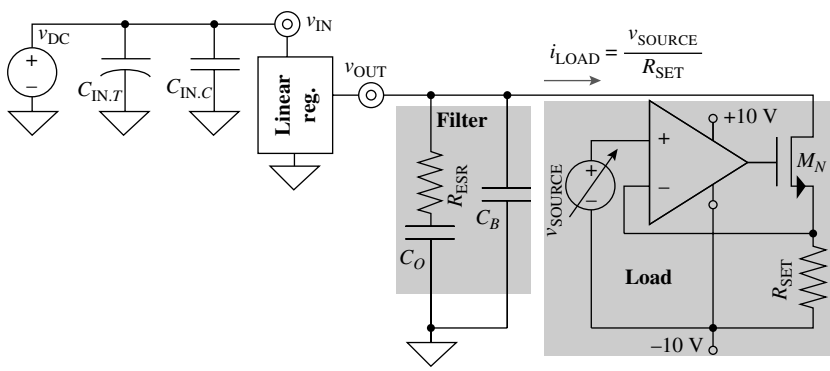


Figure 8.7 Test circuit for a linear regulator with a current-source load.

8.2.2 Regulating Performance

Load Regulation

As described in Chap. 1, load regulation (LDR) refers to how much v_{OUT} varies when subjected to full-scale, steady-state changes in i_{LOAD} (e.g., 20 mV of output-voltage variation across a load range of 0–5 mA). Extracting LDR amounts to monitoring v_{OUT} while subjecting the linear regulator, as shown in Figs. 8.6 and 8.7, to a variable resistor load R_{LOAD} and/or user-defined current-source load M_N. In these representative setups, as in most, changes in R_{LOAD} or v_{SOURCE} (i.e., load) must be slower (i.e., longer) than the speed (i.e., response time) of the regulator to ensure v_{OUT} settles to its steady-state value before recording a measurement reading. In other words, *soak time*, which ultimately translates to *test time* in production, should be long enough to capture v_{OUT} after it settles. After testing the IC at a nominal V_{IN} setting across process (i.e., dies, wafers, and lots) and temperature (e.g., –40°C, 27°C, and 125°C), a summary table reports the resulting three-sigma (i.e., 3σ) variation in V_{OUT} and an accompanying V_{OUT}-I_{LOAD} LDR graph (as shown in Fig. 8.8) illustrates a sample response.

Line Regulation

Line regulation (LNR) refers to how much v_{OUT} varies when subjected to full-scale, steady-state changes in v_{IN} (e.g., 8 mV of output-voltage variation across a line range of 1.8–3 V). The setup and procedure for extracting LNR performance closely resemble its LDR counterpart (including extending soak time beyond the response time of the circuit), except for sweeping v_{IN} and holding I_{LOAD} constant (at a nominal value). Note LNR only includes the V_{IN} range for which parametric compliance is required so it should not extend down to the circuit's headroom limit (i.e., LNR's V_{IN} range should remain above headroom limit $V_{IN(min)}$). As a result, to avoid mixing the temporary effects of a headroom crisis and start-up

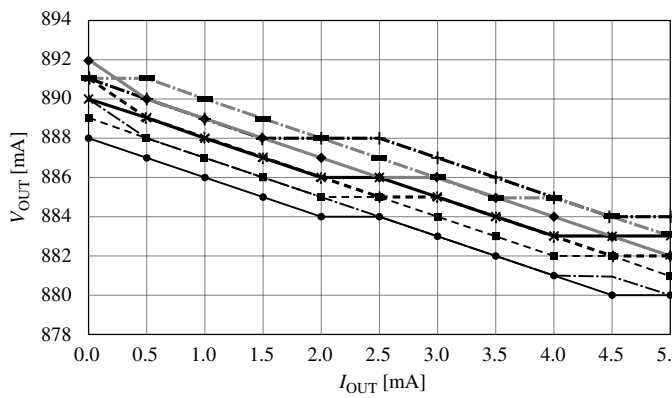


Figure 8.8 Typical load-regulation measurement results.

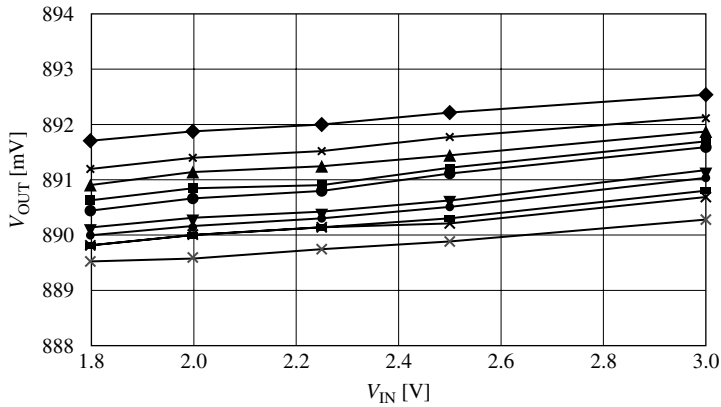


FIGURE 8.9 Typical line-regulation measurement results.

during a supply ramp-up sequence (discussed later), it is best to sweep v_{IN} from its maximum- to its minimum-specified values (e.g., from 5 to 1.5 V), monitor v_{OUT}, and discard any variations near headroom limit $V_{IN(min)}$. Again, a summary table reports the three-sigma (i.e., 3σ) variation in V_{OUT} when tested across process and temperature and an accompanying V_{OUT}-V_{IN} LNR graph (is illustrated in Fig. 8.9) demonstrates representative responses.

Temperature Drift

Temperature-drift performance is also a dc parameter and therefore requires a similar circuit and analogous measures with respect to LDR and LNR, except this time both V_{IN} and I_{LOAD} remain constant at nominal values and temperature is swept. Soak time in this measurement is normally longer because the thermal time constant of the IC in a plastic package is considerably longer than the circuit's bandwidth. If not careful with soak time, in fact, the drift in v_{OUT}, when increasing and decreasing temperature, may show false signs of thermal hysteresis (i.e., drift resulting from an ascending temperature sweep may differ from its descending counterpart). If the hysteresis is indeed an artifact of the measurement (i.e., not inherent in the reference or the regulator), which is typically the case, it should disappear when soak time extends beyond the thermal time constant of the chip.

Temperature drift in V_{OUT} across process for trimmed parts, as shown in Fig. 8.10, may not be monotonic and/or consistent, even if centered (i.e., optimized) correctly, which means quoting a slope is not entirely relevant. The best means of describing drift performance is, as in a reference, the *box method*, the goal of which is to enclose all v_{OUT} values across extreme temperature settings in a "box" and quote the 3σ variation of its vertices. In other words, the specification table should report the 3σ variations of v_{OUT} at its worst temperature settings (e.g., 15 mV of total variation between –40°C and 125°C).

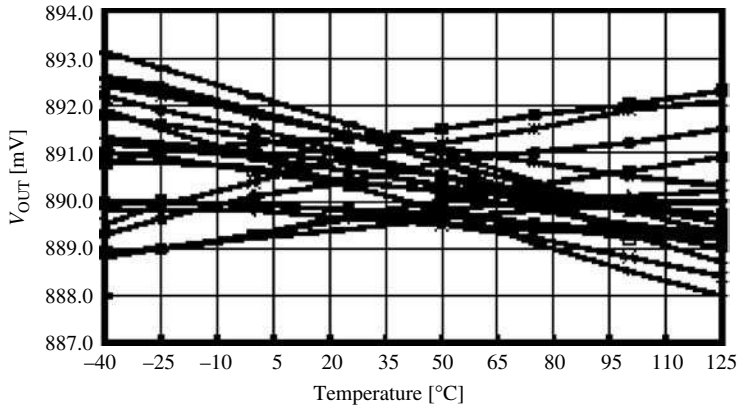


Figure 8.10 Typical temperature-drift measurement results.

Load-Dump Response

Specifying transient variations in v_{OUT} in response to sudden load changes is more difficult than its LDR, LNR, and temperature-drift predecessors are because predicting the nature of the load dumps is mostly guesswork. Even if the total transient variation in i_{LOAD} were known (and that value often differs from i_{LOAD}'s full-scale range), its rise and fall times t_r and t_f may not be so the total variation in v_{OUT} cannot be predicted because it also depends on t_r and t_f. The nominal time-domain graph the data sheets typically show is typically vague and optimistic because t_r and t_f are neither apparent from the graph nor quoted, and values used are often in the microsecond region, near the regulator's bandwidth where the circuit is fast enough to react. In all fairness, quoting difficult-to-achieve rise and fall times (e.g., 10–100 ns) may be unrealistic, especially when considering a wide application space. Nevertheless, even if not quoted, system designers understand this and normally allocate 3–7% of accuracy for load-dump-induced variations in v_{OUT}.

Unfortunately, regulator ICs targeted for SoC solutions do not enjoy the "specmanship" flexibility stand-alone devices do. The truth is modern mixed-signal systems: (1) switch at high frequencies and (2) load low on-chip output capacitances C_O, which means synchronized transitions are quick and their effects on v_{OUT} are substantial. Luckily, the load is usually more predictable and its magnitude less severe in SoC applications. Figure 8.11, just to cite a typical regulator IC example, illustrates how a 100-ns, 1–5-mA load dump at 100 kHz induces 500 mV excursions in v_{OUT}.

In practice, emulating worst-case dumps with the resistively loaded circuit shown in Fig. 8.6 by driving R_{LOAD}'s ground terminal, as shown in Fig. 8.12, with a square-wave voltage source (i.e., v_{SOURCE}) is possible. The circuit relies on the regulator's ability to fix and regulate v_{OUT} to a constant value so applying v_{SOURCE} to R_{LOAD}'s ground

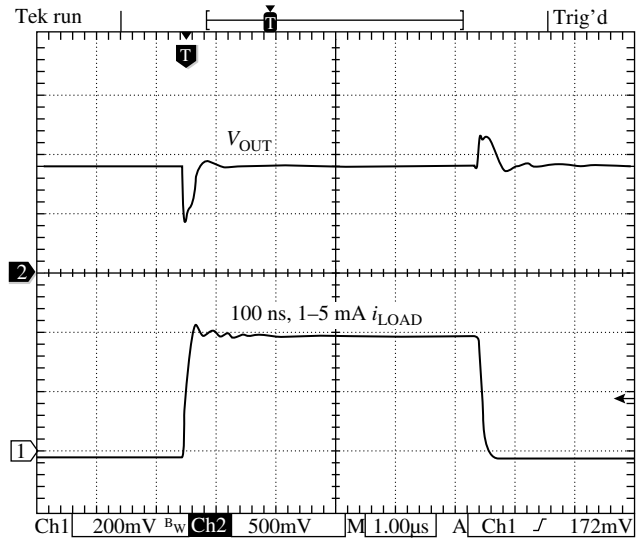


FIGURE 8.11 Sample load-dump measurement result.

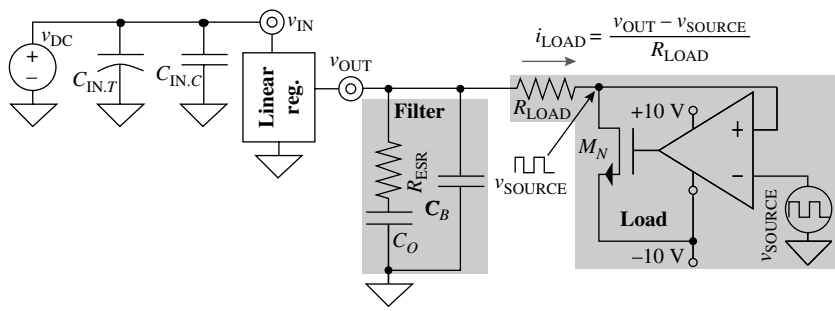


FIGURE 8.12 Modified resistively loaded test circuit for load-dump measurements.

terminal induces a square-wave load current i_{LOAD} that is equivalent to the ratio of voltage drop $v_{OUT} - v_{SOURCE}$ and R_{LOAD}. An op amp in unity-gain configuration buffers the actual signal source v_{SOURCE} because v_{SOURCE} may be incapable of sinking the desired i_{LOAD}, as is typically the case. Similarly, although not often true in SoC environments where the load is lower, the discrete op amp may not be able to sink $I_{LOAD(max)}$ so, as shown, a cascaded power NMOS transistor can be used to drive R_{LOAD}. Note the current-source circuit of Fig. 8.7 needs no modifications for this measurement, except for driving a square-wave signal through v_{SOURCE}. Also notice the op amp in all cases must be sufficiently fast and powerful to drive the power transistor's gate at the prescribed rise and fall times needed.

Accuracy

The total accuracy performance specified in a data sheet (e.g., ±1–3%) typically includes the 3σ impact of load, line, and temperature on the IC across several wafers and fabrication runs. Perhaps the most practical means of combining (and visualizing) these effects into a single specification is to monitor, measure, and graph v_{OUT} across load i_{LOAD}, as though it were an LDR measurement (as shown in Fig. 8.8), but at the extreme values of V_{IN} and temperature. In other words, the combined accuracy graph would mimic the LDR graph, except it would show four different families of curves belonging to all possible extreme combinations of V_{IN} and temperature: at the highest V_{IN} and highest temperature, lowest V_{IN} and lowest temperature, highest V_{IN} and lowest temperature, and lowest V_{IN} and highest temperature. Then, using the box method, the specification would report the 3 variations associated with the worst corners of the box. While on the subject, the accuracy specification, as quoted on a data sheet, normally excludes the effects of load dumps.

Power-Supply Rejection

One of the reasons why power-supply ripple rejection (PSRR), as is often called, also describes power-supply rejection (PSR) is because the test setup used to measure PSR literally injects a ripple into the circuit through v_{IN}. The measurement therefore consists of introducing a ripple-emulating sinusoid at a particular frequency f_{IN} into v_{IN}, monitoring the resulting ripple in v_{OUT}, measuring the ratio of the peaks in v_{IN} and v_{OUT}, and repeating the procedure at other frequencies. A graph of ripple-rejection ratio $\Delta v_{IN}/\Delta v_{OUT}$ or supply gain $\Delta v_{OUT}/\Delta v_{IN}$ (i.e., A_{IN}) across frequency, as shown in Fig. 8.13*a* where the responses of several different regulator architectures are included, would then describe PSR performance. Similarly, a list of data points would also describe PSR (e.g., 50 dB rejection at 1 kHz with a 10-mA load, etc.).

In practice, a function generator alone is not a suitable replacement for v_{IN} because the generator is normally incapable of driving the currents a linear regulator conducts. Alternatively, an op amp and a power NMOS transistor in unity-gain feedback configuration, as exemplified by the PSR circuit shown in Fig. 8.13*b*, can buffer and unload the function generator (i.e., v_{SOURCE}). Incidentally, the peak-to-peak value of the input ripple is normally on the order of 20–100 mV to mimic the ripple the output of a switching converter would inject into the circuit in a typical real-life application. The measured results, as it turns out, may not correlate exactly with the small-signal response obtained from ac simulations because 20–100 mV variations in v_{IN} may not constitute the small signals the simulator expects, especially when the regulator is close to dropout. Transient, time-domain simulations that inject a ripple into v_{IN} and monitor the resulting ripple in the output, as expected in real life and applied to the actual measurement, would more accurately predict PSR performance.

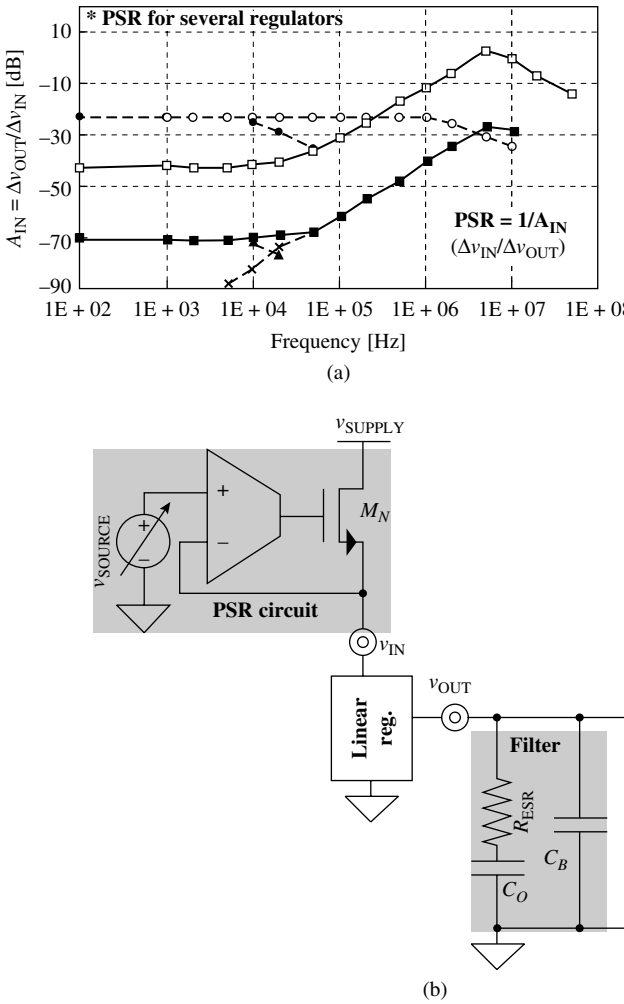


FIGURE 8.13 (*a*) Supply gain A_{IN} and PSR^{-1} response for different linear regulator topologies and (*b*) the circuit used to generate them.

8.2.3 Power Performance

Efficiency

Determining the efficiency performance of a linear regulator reduces to measuring input current I_{IN} across the entire load-current range because, outside of I_{IN} and I_{LOAD}, efficiency η only depends on input and output voltages V_{IN} and V_{OUT}, both of which the user or target application set:

$$\eta = \frac{P_{OUT}}{P_{IN}} = \frac{I_{LOAD} V_{OUT}}{I_{IN} V_{IN}} = \eta_I \left(\frac{V_{OUT}}{V_{IN}} \right) \tag{8.8}$$

As such, because V_{IN}, V_{OUT}, and I_{LOAD} are normally outside the control of the IC designer (i.e., they are open system-design variables), extracting and quoting ground current I_{GND} (or equivalently, quiescent current I_Q) and/or current efficiency η_I are often more meaningful, where

$$\eta_I \equiv \frac{I_{LOAD}}{I_{IN}} = \frac{I_{LOAD}}{I_{GND} + I_{LOAD}} \tag{8.9}$$

and I_{GND} is extracted from I_{IN} by subtracting I_{LOAD} from I_{IN} during an LDR measurement. A summary table ultimately specifies 3σ variations in I_Q (or equivalently, I_{GND}) at the extremes of the load range, or a nominal value at the halfway point (e.g., 20 μA of I_Q at a load of 10 mA). I_{GND} (or I_Q) is not always a function of I_{LOAD}, but if it were, a nominal graph of I_{GND} with respect to I_{LOAD} may also accompany the specification table.

Dropout

Measuring dropout voltage V_{DO} is not always straightforward. The difficulty mostly arises in applications that demand low output voltages V_{OUT} because decreasing V_{IN} may push the regulator beyond its headroom limit $V_{IN(min)}$ before allowing it to enter the dropout region. In those cases, increasing V_{OUT}'s target regulation point (by modifying the resistor feedback network) increases the input onset voltage of the dropout region so that decreasing V_{IN} will first push the regulator into dropout before letting headroom run its course. To secure a valid measurement, as a result, understanding how the regulator behaves in all its various regions of operation is important.

Consider the results shown in Fig. 8.14, as obtained by expanding the LNR measurement range to include lower V_{IN} values. In decreasing

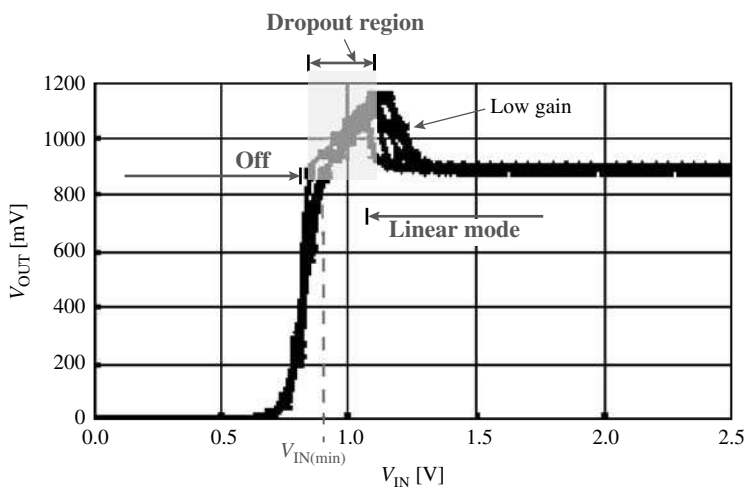


Figure 8.14 Sample dropout and headroom limit $V_{IN(min)}$ measurement results from an expanded LNR measurement.

V_{IN} down from 2.5 to 1.25 V, the linear regulator tested remains in the linear region, regulating v_{OUT} to its target of 0.9 V. Between 1.15 and 1.25 V, V_{IN} pushes one or several transistors in the regulator closer to triode so the regulator's loop gain decreases and v_{OUT} noticeably deviates from its target. (Note v_{OUT} may not always increase with decreasing values of V_{IN} in the low-gain portion of the linear region.) In decreasing V_{IN} beyond this point, below 1.15 V for example, once there is no more loop gain, the regulator enters dropout and v_{OUT} decreases almost linearly with V_{IN}. In this region, V_{DO} is $V_{IN} - V_{OUT}$, except data sheets normally report V_{DO} only at maximum load current $I_{LOAD(max)}$, at the worst possible load setting (e.g., 50 mV drop at 10 mA). Upon reaching $V_{IN(min)}$, 0.9 V in this case, the regulator ceases to operate properly so the circuit no longer closes the negative-feedback loop required to regulate v_{OUT} and I_{LOAD} consequently discharges C_{OUT} and pulls v_{OUT} to 0 V.

8.2.4 Operating Environment

Load Range

Although not always the case, the onset of dropout sometimes sets the regulator's maximum load-current limit $I_{LOAD(max)}$. It is not unusual for industry, for instance, to determine the current necessary to induce a dropout voltage of 200 mV and use that setting as $I_{LOAD(max)}$. Consider, for example, setting V_{IN} to maybe 100mV above the regulated output, sweeping I_{LOAD}, and monitoring which load setting induces v_{OUT} to drop 200 mV below V_{IN} determines the maximum load-current limit (i.e., $I_{LOAD(max)}$) that produces a 200-mV dropout. This method of extracting $I_{LOAD(max)}$, however, only applies to stand-alone regulator chips whose target design space is not as clear as their application-specific counterparts. System-on-chip (SoC) designs, just to cite an increasingly ubiquitous example, must supply specific load-current ranges under predetermined dropout conditions, which means determining $I_{LOAD(max)}$ is not as relevant or meaningful as extracting LDR performance across the desired load range.

Headroom

As discussed and illustrated earlier in Fig. 8.14, an expanded LNR measurement also includes the effects of headroom. In more explicit terms, as V_{IN} decreases, v_{OUT} drops abruptly at headroom limit $V_{IN(min)}$ (e.g., at 0.9 V in Fig. 8.14). Note again that increasing V_{IN} from ground may not exhibit the same results because v_{OUT} may otherwise also include the effects of start-up and power-up sequences.

Output Filter

Determining the acceptable filter space for which the regulator is stable is somewhat complicated, when compared to other measurements. The main problem, as before, is the unpredictable nature of the load, as overstressing the regulator with excessively fast and large

load dumps seems pessimistic and understressing it seems unrealistic. Coupling this uncertainty with how involved the measurement is in practice explains the inconsistency of the methods used to extract this range in industry and the seemingly incomplete data sheets that result. Consider, however, that stability (as opposed to phase margin, which implies proneness to unstable conditions) is ultimately a binary parameter because the circuit under a given filter scenario is either stable or unstable. As a result, stability tests should include all possible worst-case conditions such as full-scale load dumps (e.g., from 0 A to $I_{LOAD(max)}$) with the fastest rise and fall times possible (e.g., less than 10–100 ns at maybe 100 Hz) at the extreme limits of the temperature range. (Note slower rise and fall times do not expose the circuit to high-frequency perturbations.)

The difficulty in this test is implementing output capacitor-ESR combination C_O-R_{ESR}. The easiest approach, although not always the most practical, is to find enough capacitors to populate the entire C_O-R_{ESR} space, use them to filter v_{OUT} under fast load-dump conditions, and monitor if oscillations result— the circuit is stable if no C_O-R_{ESR} setting produces oscillations. The problem with this method is finding sufficient capacitors to characterize the regulator fully. Bear in mind that using transistors to switch resistors in and out of an artificial R_{ESR}-resistor string in combination with a low-ESR capacitor is not practical because the transistor's switch resistance is on the order of the R_{ESR} targeted, which means the transistor constitutes a considerable parasitic.

Alternatively, as shown in Fig. 8.15, soldering several low-ESR ceramic multilayer 1–10 nF capacitors in parallel and the entire array in series with several 1 Ω power resistors (in parallel) to build and emulate several C_O-R_{ESR} combinations is viable. With this tactic, the first measurement setting corresponds to the highest possible C_O and lowest possible R_{ESR} combination, when all capacitors and resistors are used. The procedure continues by disconnecting (i.e., desoldering) one resistor from the array at a time, after monitoring the response of v_{OUT} to load dumps at both low and high temperatures. The next set of measurements involve

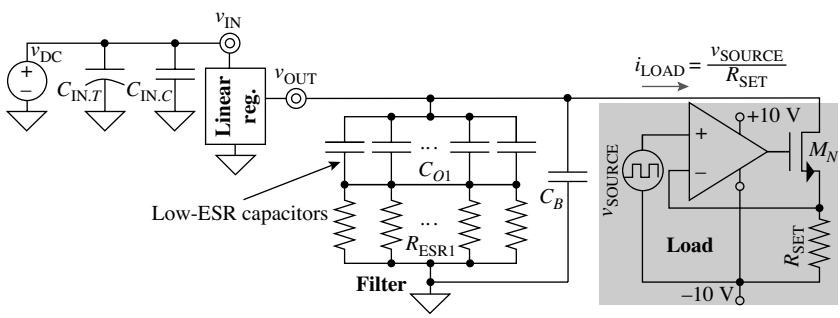


Figure 8.15 Test circuit used to test the C_O-R_{ESR} stability space of a linear regulator.

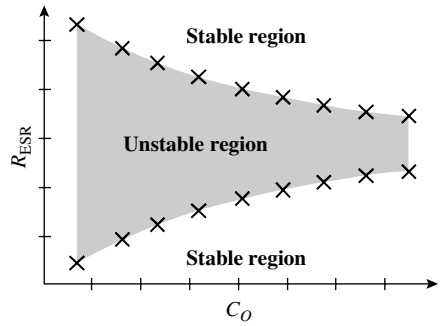


FIGURE 8.16 Representative C_O-R_{ESR} stability space graph corresponding to an externally compensated regulator showing the so called "curves of death."

reconnecting (i.e., resoldering) all resistors, disconnecting (i.e., desoldering) another capacitor, and repeating the ESR-reduction process. Resistors are once again reconnected (i.e., resoldered), another capacitor is disconnected (i.e., desoldered), and the ESR-reduction process is repeated until the entire C_O-R_{ESR} space is tested.

Figure 8.16 shows the representative C_O-R_{ESR} space that results when subjecting an externally compensated regulator to stability measurements. Each data point represents the edge beyond which an increase or decrease in R_{ESR} causes sustained oscillations in v_{OUT}. Some designers call the C_O-R_{ESR} boundaries that result "the curves of death" because they map where the regulator is and is not stable, in other words, the C_O-R_{ESR} combinations for which the circuit survives and recovers from high-frequency perturbations. As expected, the graph shown depicts a regulator that is generally more stable with increasing output capacitance C_O, which corresponds to a regulator whose output pole p_O is dominant. For stable conditions to persist at low C_O values, R_{ESR} must either be low enough to avoid the zero R_{ESR} controls from extending unity-gain frequency f_{0dB} into the parasitic-pole region or, conversely, sufficiently high for the ESR zero to cancel the effects of the system's second dominant pole, which is, for the most part, amplifier pole p_A. An internally compensated response generally complements the graph shown because the regulator would generally become more stable with decreasing C_O values. With high output capacitances, an internally compensated regulator normally requires sufficiently high R_{ESR}'s for the ESR zero to offset the effects of p_O or low enough R_{ESR}'s to avoid pulling the ESR zero below f_{0dB}, where another zero may already exist for the purpose of canceling p_O.

8.2.5 Start-Up

Start-up or *power-up* sequences are often unspecified because the mere fact the data sheet or the designer claims the circuit works is a guarantee. Emerging portable applications, however, are starting to demand this specification not because they wish to verify its functionality but because they can only tolerate short start-up times (e.g., start in less than 1 μs). The driving motivation for characterizing this

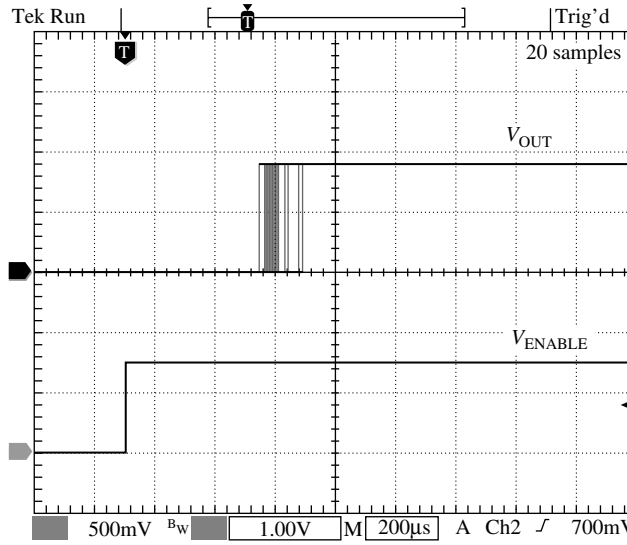


FIGURE 8.17 Superimposed enabled start-up measurements.

parameter is that small, battery-powered devices cannot afford to dissipate any quiescent power indefinitely and must therefore engage and disengage circuits to preserve energy and extend operational life. As a result, testing start-up is not only a functional state test but also a quantitative measurement.

Start-up tests, as in a real-life application, must consider supply ramps, supply and load pulses, and enabled starts. Since full supply ramps can occur in both micro- and milliseconds, tests should include both extreme sets of conditions. Intermediate rise times should also be included to ensure the corresponding frequency of the supply ramp does not happen to prompt uncontrolled oscillations in v_{OUT}. Similarly, the regulator should recover from noise- and load-induced pulses in v_{IN} (e.g., 2-V pulses with a duration of 1 µs) and v_{OUT} (e.g., 50-mA load pulses with a duration of 1 µs), the latter of which is already tested in the load-dump measurement. To this end, the same circuit used to test PSR (as shown in Fig. 8.13*b*) is a good setup for all ramped and pulsed v_{IN} measurements. In the case of enabled regulators, response times to enable ramp-up sequences with less than 100 ns rise times are quantified (e.g., the IC starts after 500 µs of being enabled) and graphed, as illustrated in Fig. 8.17.

8.3 Summary

Protecting a linear regulator IC is as important as designing it in the first place. Consider that, in practice, just to cite an example, the load may short and stress the IC with substantially high currents, which

when sustained, can elevate the IC's junction temperature beyond the melting point of the plastic package. Including overcurrent protection and/or thermal-shutdown features in the IC, as a result, safeguard the regulator, and more specifically, the power device from such strenuous and damaging conditions. The designer must also anticipate the human factor because people (inadvertently or not) may not only reverse the battery but also expose the chip to sudden electrostatic discharges. Irrespective of the means and shielding objectives, one of the most important challenges in including protection and staying alert for all possible electrical violations is maintaining transparency, which is to say the protection circuit should neither dissipate power nor introduce parasitic voltage drops.

Ignorance on the part of the user or system designer can be equally destructive, if not more, because systematically overstressing the regulator with a load it cannot handle may not fail immediately but later, when a product using the IC is in full production. Characterization, in this regard, is by far the best means of increasing reliability and mitigating the risk of a "call back." The data, however, is only as valid as the measurements that extract them, which is why emulating the load and carefully considering the test setup are important in this process. While a resistor alone, for example, may fully model a particular load, it cannot represent all possible loads, so complementing the measurements with a current-source load is often a sensible precaution. Bear in mind that, although not always tangibly apparent or evident, the savings from avoiding short- and long-term product failures justifies (within reason) the test time and effort (i.e., cost) dedicated to fully ascertaining the regulation, power, and operating limits of the regulator.