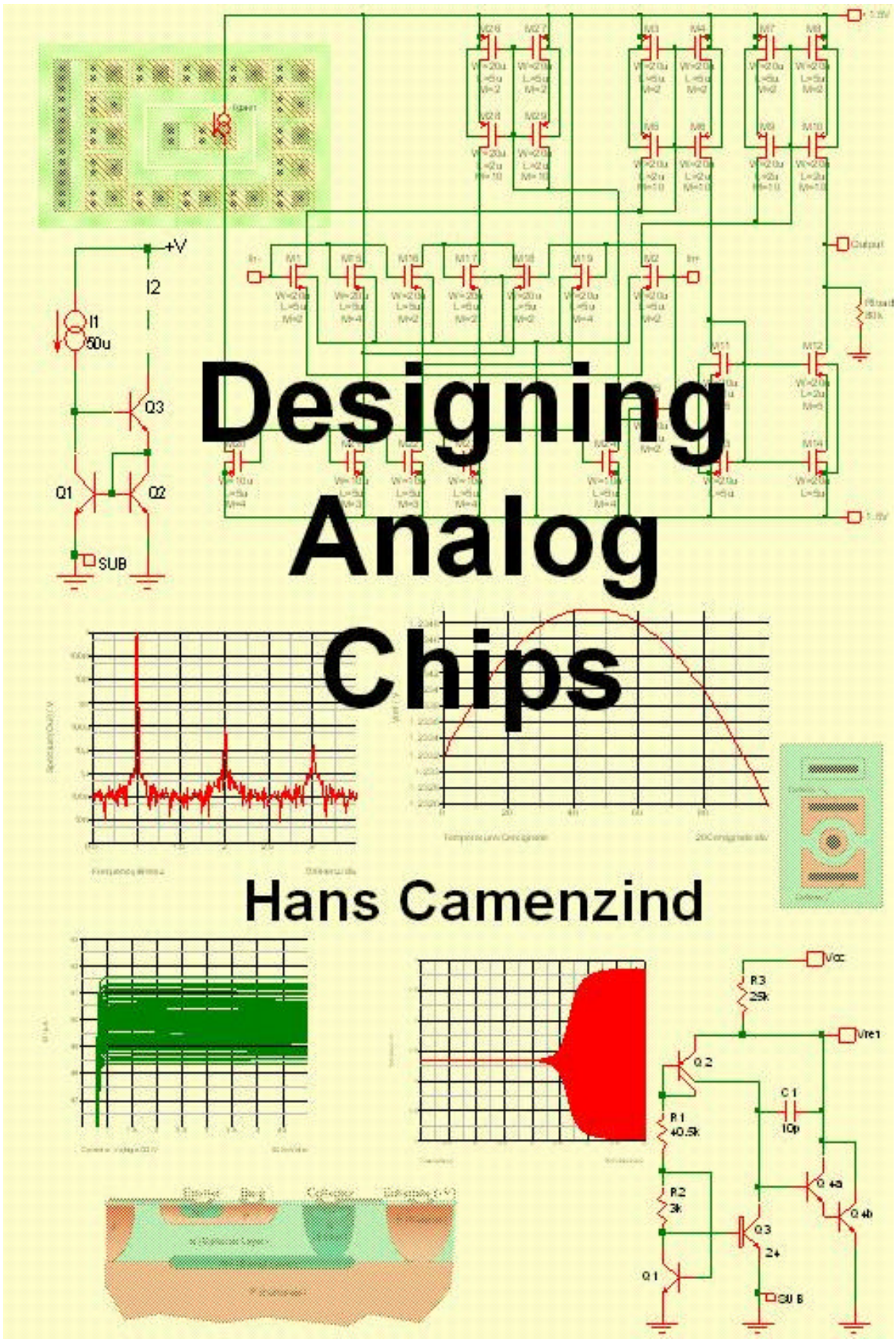

+V
I2
I1
50u
Q3
Q1
Q2
SUB
Designing
Analog
Chips
Hans Camenzind
Vcc
R3
25k
Vref
Q2
C1
10p
R1
40.5k
Q4a
Q4b
R2
3k
Q3
24
Q1
SUB

(camenzind@arraydesign.com)

February 2005

The author is indebted to the following for comments, suggestions and corrections:

Bob Pease, Jim Feit, Ted Bee, Jon Fischer, Tim Camenzind, Jules Jelinek, Brian Attwood, Ray Futrell, Beat Seeholzer, David Skurnik, Barry Schwartz, Dale Rebgetz, Tim Herklots, Jerry Gray, Paul Chic, Mark Leonard, Yut Chow, Gregory Weselak, Lars Jespersen and Wolfgang Horn.

Table of Contents

Analog World

"Everything is going digital". Cell phones, television, video disks, hearing aids, motor controls, audio amplifiers, toys, printers, what have you.

Analog design is obsolete, or will be shortly. Or so most people think.

Imminent death has been predicted for analog since the advent of the PC. But it is still here; in fact, analog ICs have been growing at almost exactly the same rate as digital ones. A digital video disk player has more analog content than the (analog) VCR ever did.

The explanation is rather simple: the world is fundamentally analog. Hearing is analog. Vision, taste, touch, smell, analog all. So is lifting and walking. Generators, motors, loud-speakers, microphones, solenoids, batteries, antennas, lamps, LEDs, laser diodes, sensors are fundamentally analog components.

The digital revolution is constructed on top of an analog reality. This fact simply won't go away. Somewhere, somehow you have to get into and out of the digital system and connect to the real world.

Unfortunately, the predominance and glamour of digital has done harm to analog. Too few analog designers are being educated, creating a void. This leaves decisions affecting analog performance to engineers with a primarily digital background.

In integrated circuits, the relentless pressure toward faster digital speed has resulted in ever-decreasing supply voltages, which are anathema to high-performance analog design. At 350nm (3.3V) there is still enough headroom for a high-performance analog design, though 5 Volts would be better. At 180nm (1.8V) the job becomes elaborate and time-consuming and performance starts to suffer. At 120nm (1.2V) analog design becomes very difficult even with reduced performance. At 90nm, analog design is all but impossible.

There are "mixed signal" processes which purportedly allow digital and analog circuitry on the same chip. A 180nm process, for example, will have some devices which can work with a higher supply voltage (e.g. 3 Volts). While such an addition is welcome (if marginal), the design data (i.e. models) are often inadequate and oriented toward digital design.

Hence this book. It should give you an overview of the world of analog IC design, so that you can decide what kind of analog function can and cannot, should and should not be integrated. What should be on the same chip with digital and what should be separate. And, equally important, this book should enable you to ask the right questions of the foundry, so that your design works. The first time.

* * *

You will find that almost all analog ICs contain a number of recognizable **circuit elements**, functional blocks with just a few transistors. These elements have proven useful and thus re-appear in design after design. Thus it makes sense to first look at such things as current mirrors, compound transistors, differential stages, cascodes, active loads, Darlington connections or current sources in some detail and then examine how they are best put together to form whole functions.

* * *

Academic text books on IC design are often filled with mathematics. It is important to understand the fundamentals, but it is a waste of time to calculate every detail of a design. Let the simulator do this chore, it can do it better and faster than any human being. An analysis will tell you within seconds if you are on the right track and how well your circuit performs. Assuming that you have competent models and a capable simulator, an analysis can teach you more about devices and circuits than words and diagrams on a page.

1 Devices

Let's assume your IC design needs an operational amplifier. Which one? If you check the data-books of linear IC suppliers, you'll find hundreds of them. Some have low current consumption, but are slow. Others are quite complex, but feature rail-to-rail inputs and outputs. There are inputs which are factory-trimmed for low offset voltages, outputs for high currents, designs for a single supply voltage, very fast devices, etc.

Here is the inherent problem with analog building blocks: there are no ideal designs, just configurations which can be optimized for a particular application. If you envisioned a library from which you can pull various analog building blocks and insert them into your design, you are about to experience a rude shock: this library would have to be very large, containing just about every operational amplifier (and all other linear functions) listed in the various data-books. If it doesn't, your IC design is bound to be inferior to one done with individual ICs.

In short: *There are no standard analog cells*. If your application is the least bit demanding, you find yourself either modifying previously used blocks or designing new ones. In either case you need to work on the device level, connecting together transistors, resistors and rather small capacitors.

To do this you need to know what devices are available and what their limitations are. But above all you need to *understand* devices in some detail. The easiest way to learn about complex technical things is to follow their discovery, to have the knowledge gained by the earlier men and women (who pioneered the field) unfold in the same way they brought it to light.

Semiconductors

In 1874 Ferdinand Braun was a 24-year old teacher in Leipzig, Germany. He published a paper which was nothing short of revolutionary: he had found that some materials violated Ohm's law. Using naturally formed crystals of Galena (lead sulfite, the chief ore mineral of lead) and other sulfites, he pressed a spring-loaded metal tip against their surfaces and observed that the current through this arrangement was dependent on the polarity of the applied voltage. Even more puzzling was the fact that, in the direction which had better conduction, the resistance decreased as the current was increased.

What Braun (who later would give us the CRT) had discovered, we now know as the diode, or rectifier. It was not a very good one, there was only a 30% difference between forward and reverse current. And there were no practical applications. Braun could not explain the effect, nor could anybody else.

In 1879 Edwin Hall of Johns Hopkins University discovered what was later named the Hall Effect: when you pass a magnetic field through a piece of metal it deflects the current running through the metal. In all the metals he tried the deflection was to one side; he was greatly relieved to see that this confirmed the negative charge on the electron.

But then the surprise came. In some materials the deflection went the other way. Where there perhaps positive electrons?

Nothing much happened until about 1904. Radio appeared on the scene and needed a "detector". The signal was amplitude modulated and to make the music or speech audible the radio frequency needed to be rectified (i.e. averaged). Thus, 30 years after Braun's discovery, the "odd behavior" of a wire touching Galena (and now many other materials, such as silicon carbide, tellurium and silicon) found a practical application. The device was called the "Cat's whisker", but it actually didn't work very well; one had to try several spots on the crystal until one was found which produced a loud enough signal.

And it was replaced almost immediately by the vacuum tube, which could not only rectify but amplify as well. Thus the semiconductor rectifier (or diode) went out of fashion.

It was not until 1927 that another practical application appeared: large-area rectifiers. These were messy, bulky contraptions using copper-oxide (and later selenium) to produce DC from line voltage, chiefly to charge car batteries. But there was still no understanding of how these devices worked.

In the background, mostly at universities and large corporate laboratories, some research went on, despite the fact that there was no semiconductor industry yet. In 1931 A.H. Wilson came up with a complete model of energy bands: electrons exist only at discrete levels, each with a higher energy than the lower one; only two electrons can exist at the same level, but they have opposite spins; at the last (or highest level) are the valence electrons and there is a gap in energy to the ultimate one, the conduction band; once they reached that last level, conduction happens by accelerating the electrons in an electric field.

The theory was fine, but it took 15 years for someone to make a connection between it and the diode.

There were two problems masking the real semiconductor effects. First, all the behaviors so far noticed were surface effects. The cat's whisker applied a metal wire, the copper-oxide and selenium rectifier metal plates. Today this is recognized as a rather specialized configuration, only surviving in the Schottky diode. Second, the semiconductor material was anything but pure, containing elements and molecules which counteracted the desired behavior.

Then World-War II happened and with it came radar. To get adequate resolution, radar needed to operate at high frequencies. Vacuum tubes were too slow, so the discarded "cat's whisker" came into focus again (employed right after the antenna to rectify the wave so it could be mixed with a local oscillator and produce a lower frequency, which could be handled by vacuum tubes).

This time a world-wide emergency drove the effort, with plenty of funding for several teams. They started with the "cat's whisker" and tried to determine what made it so fickle and unreliable. It became immediately obvious that purer material was required, and that this material should be in the form of a single crystal. When they heated part of a crystal close to the melting point and moved the heated zone, the foreign materials moved with it. And now they realized that some of these impurities were actually *required* to get the diode effect. And these impurities all fell into very specific places within the periodic table of elements.

Silicon and **germanium** both have a valence of four. Valence simply means that in the outermost layer of electron orbits there are four electrons. Silicon, for example, is element number 14, meaning it has a total of 14 electrons. The first orbit (or energy level) has two electrons, the second eight and the third four.

The outermost orbits of the atoms touch each other and the electrons in this orbit don't stay with one particular atom, they move from orbit to orbit. It is this sharing of electrons that hold the atoms together. And this

ability to move from atom to atom is also the basis of electrical conduction: in conductors the electrons roam widely and are easily enticed to move in an electrical field, whereas in an insulator they stay close to home.

Electrically, pure silicon is a terribly uninteresting material. It is an insulator, but not a very good one. The fun begins when we add the right impurities, or dopants.

Just to the right of silicon in the periodic table is **phosphorus**, element number 15. Like Silicon, it has two electrons in the first orbit, eight in the second but there are five in the third. Now let's say we were able to pluck out an atom in a block of silicon and replace it with a phosphorus atom. Four of the valence electrons of this new atom will circulate with the silicon electrons, but the fifth one won't fit in. This excess electron creates a negative charge and the silicon becomes what we now call n-type.

This introduction of excess electrons is unlike static charge. When you brush your hair so that it stands upright, you have simply moved some electrons temporarily. When you "dope" silicon, the charge is permanent, fixed in the crystal lattice (and does not become a battery).

Similarly, to the left of Silicon and one space up in the periodic table is **boron**, element number 5. It has two electrons in a first level and three in a second, a valence of three. If we replace a silicon atom with a boron one, there is an electron missing and we create a positive charge, or p-type material. As with the excess electron in n-type silicon, we can apply an electric field and cause a current to flow, but the net-effect is the flow of holes, not electrons. This is what makes the Hall effect go the wrong way.

It is important to understand this mechanism of moving holes and electrons in doped semiconductors. In n-type material an excess phosphorus electron wanders into the path of a neighboring silicon electron and displaces it. The displaced electron then takes the orbit of another one and so on until the last electron ends up at the starting point, the phosphorus atom.

This endless game of musical chairs - proceeding at near the speed of light - depends greatly on the temperature. At absolute zero there is no movement. At about -60^{o}C the movement is sufficient for semiconductor effect to start in silicon. At about 200^{o}C there is so much movement that silicon practically becomes a conductor. It is only within a relatively narrow range, about -55^{o}C to 150^{o}C, that silicon is a useful semiconductor.

In p-type material the movement starts with an electron in the neighborhood of the boron atom. It fills the vacancy and then is itself replaced by another electron and so on until the first electron moves away

from the boron atom again. The moving is done by electrons, but the net effect is a moving hole.

When an electric field is present the movement takes on a direction: electrons flow toward the positive electrode and are replaced by other electrons flowing out of the negative electrode.

It is amazing how few dopants it takes to make n-type or p-type material. Silicon has $5x10^{22}$ atoms per cubic centimeter. A doping level can easily be as low as $5x10^{15}$ boron or phosphorus atoms per cubic centimeter, i.e. one dopant atom for every 10 million silicon atoms. No wonder it took so long to discover the true nature of the semiconductor effects; in nature, the number of miscellaneous impurities is far larger than one in 10 million.

The Diode

Even with a dopant present silicon is uninteresting. It is not a good conductor and as a resistor it is inferior to metal film or even carbon. But if we have both n-type and p-type atoms in the same silicon crystal, things suddenly happen.

Opposite charges attract each other, so the excess electrons near the border of the n-type section move into the p-type material and stay there. An electron fills a hole and the electric charges cancel each other.

This only happens over a short distance, as far as an electron (or hole) can roam. The resulting region is called the space-charge layer or *depletion region*.

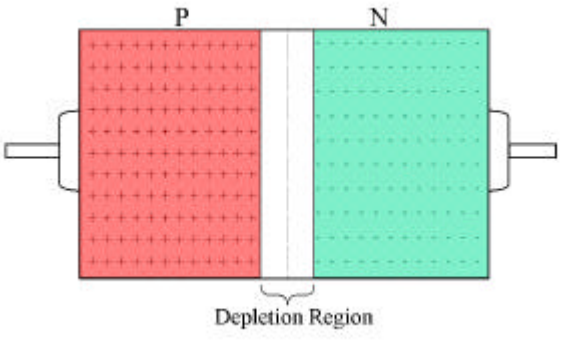


Fig. 1-1: A depletion region forms between p-doped and n-doped semiconductor areas.

Now suppose you connect a voltage to the two terminals. If the p-region is connected to the negative terminal of the supply and the n-region to the positive one, you simply push the charges away from each other, enlarging the depletion region.

If, however, the p-region is positive and the n-region negative, you push the charges closer together as the voltage increases. The closer proximity forces more and electrons and holes to

cross the depletion region. The effect is exponential: at 0.3 Volts (at room temperature) very little current flows; at 0.6 Volts the current is substantial and at 0.9 Volts very large.

The expression for the diode voltage is:

$$Vd = \frac{kT}{q} \ln \frac{I1}{Is} \qquad \text{or} \qquad I1 = Is\left(e^{\frac{Vd \cdot q}{kT}} - 1 \right)$$

where Vd = voltage across the diode
k = Boltzman constant (1.38E-23 Joules/Kelvin)
T = the absolute temperature in Kelvin
q = the electron charge (1.6E-19 Coulombs)
I1 = the actual current through the diode
and Is = diffusion current

Note that 1.38E-23 is a more convenient notation for 1.38×10^{-23}.

The diffusion current Is depends on the doping level of n-type and p-type impurities, the area of the diode and (to a very high degree) on temperature. A reasonable starting point for a small-geometry IC diode is Is=1E-16.

The equations neglect a few things. There is a limit in the voltage that can be applied in the reverse direction. Similar to an arc-over in any insulator, there comes a point when the electric field becomes too large and the opposing charges crash into each other. This *breakdown voltage* depends on the concentration of dopants: the higher the concentration, the lower the breakdown voltage.

There is a price to be paid for high breakdown voltage. As the dopant concentration is lowered, the depletion layer becomes larger and the higher voltage pushes it deeper yet. This distance must be accommodated in the design.

The opposing charges in a semiconductor junction are no different from those on the plates of a capacitor. So every junction has a capacitance; but since the distance between the electrons and holes changes with applied voltage, the capacitance becomes voltage dependent. The lower the voltage, the higher the capacitance, increasing right into the forward direction.

Lastly, there is resistance in the semiconductor material not taken up by the depletion region. For our "typical" concentration of 5E15 (atoms per cubic centimeter, giving a practical breakdown voltage in an IC of about 25 Volts), the resistivity is about 1 Ohm-cm for phosphorus (n-type) and 3 Ohm-cm for boron (p-type). For comparison, aluminum has a resistivity of 2.8 microOhm-cm, copper 1.7 microOhm-cm. Resistivity (ρ or rho) is

measured between opposite surfaces of a cube of material with a side-length (w, h, l) of 1cm (10mm):

$$r = \frac{R * w * h}{l} = \frac{Ohm * cm * cm}{cm} = Ohm * cm \text{ (or Ohm-cm)}$$

The (Bipolar) Transistor

At the time of the first serious work on the semiconductor diode, Bell Laboratories in New Jersey was already world-famous. It attracted the brightest scientists and, even among those, Bill Shockley was a stand-out. In 1938 Shockley teamed up with Walter Brattain to investigate semiconductors.

The depletion layer intrigued Shockley. There was a faint similarity to the vacuum diode. It occurred to Shockley that, if he could somehow insert a grid into this region, it might be possible to control the amount of current flowing in a copper-oxide rectifier, creating the solid-state equivalent of the vacuum triode. Shockley went to Brattain with the idea and Brattain was amused. The same idea had occurred to him too; he had even calculated the dimensions for such a grid, which turned out to be impractically small. Shockley tried it anyway and couldn't make it work. Brattain had been right.

Shockley was not a man easily defeated, though. He modified his idea and came up with a different principle of operation. He conceived that, since a relatively small number of electrons or holes are responsible for conduction in semiconductors and they each carry a charge, he could place a metal electrode near the surface, connect it to a voltage and thus either pull these carriers toward the surface or push them away from it. Therefore, he thought, the conduction of the region nearest the surface could be altered at will. He tried it -- and it didn't work either. The idea was identical to today's MOS transistor.

The work stopped there; both Shockley and Brattain were assigned to other projects during the war. But in 1945 Shockley was made co-supervisor of a solid-state physics group which included Brattain. Shockley was 35, Brattain 43. The progress made in refining silicon and germanium was not lost on Shockley; he decided to try his idea for an amplifying device again and had a thin film of silicon deposited, topped with an insulated control electrode. It still didn't work; no matter what voltage was applied to the control electrode, there was no discernable change in current through the silicon film. Shockley was puzzled; according to his

calculations there should have been a large change. But the effect - if there was any - was at least 1500 times smaller than theoretically predicted.

It was at this time, that John Bardeen, 37, joined Shockley's group. He looked at Shockley's failed experiment and mulled it over in his head for a few months. In March 1946 he came up with an explanation: it was the surface of the silicon which killed the effect. Where the silicon stops, the four valence electrons are no longer neatly tied up by the neighboring atoms. Bardeen correctly perceived that some of them were left dangling and thus produced a surface charge (or voltage), which blocked any voltage applied to an external control electrode.

With this theoretical breakthrough the group now decided to change directions; instead of attempting to make a device, they investigated the fundamentals of semiconductor surfaces. It was a long, painstaking investigation; it took more than a year. On November 17, 1947 Robert B. Gibney, another member of the group and a physical chemist, suggested using an electrolyte to counteract the surface charge. On November 20 he and Brattain wrote a patent disclosure for an amplifying device as tried by Shockley but using electrolyte on the surface. Then they went to the lab and made one. The electrolyte was extracted from an electrolytic capacitor with a hammer and nail. The device worked, the electrolyte did precisely the job that Gibney thought it would.

But, although this "field effect" device amplified, it was very slow, amplifying nothing faster than about 8Hz. Brattain and Bardeen suspected that it was the electrolyte that slowed down the device so, on December 16, 1947, they tried a different approach: a gold spot with a small hole in the center was evaporated onto germanium, on top of the insulating oxide. The idea was to place a sharp point-contact in the center without touching the gold ring, so that the point would make contact with the germanium, while the insulated gold ring would shield the surface. And now, for the first time, they got amplification.

There was only one thing wrong with this device: it didn't work as expected. A positive voltage at the control terminal increased the current through the device when, according to their theory, it should have decreased it. Bardeen and Brattain investigated and found they had inadvertently washed off the oxide before evaporating the gold, so that the gold was in contact with the germanium. What they were observing was an entirely different effect, an injection of carriers by the point contact. They realized that, to make such a device efficient, the distance between the two contacts at the surface needed to be very small. They evaporated a new gold spot, split it in half with a razorblade and placed two point contacts on top. Now the device worked even better and they demonstrated it to the Bell

management on December 23, 1947.

For half a year Bell kept the breakthrough a secret. Bardeen and Brattain published a paper on June 25. 1948 and on June 30 a press conference was held in New York. The announcement made little impression; the New York Times devoted a few lines to it on page 46.

Shockley had been disappointed by the turn of events, he had not been part of the final breakthrough. But he realized that, even though there was a working device, the battle wasn't over yet. No-one within the group really understood precisely how the transistor worked. So, in the early days of January 1948 Shockley sat down and tried to figure out what was going on between the two point contacts. And in the process he conceived a much better structure: the junction transistor.

It was a brilliant analysis and holds up to this day. In a bipolar transistor there is a current flowing between the base and emitter terminals, which is a diode. Thus electrons flow from the emitter to the base (so named because in the original point-contact transistor it was the bulk of the material). Since the base is p-doped, these electrons are the *minority carriers* in the base (hence the name bipolar transistor - carriers of both polarities are needed for the effect). A few of them will reach the base terminal. But if the base is lightly doped and *very thin* most of them will be attracted by the positive collector voltage before they re-combine with a hole in the base. In a good transistor 100 or even 500 of the electrons will be side-tracked to the collector while one goes to the base terminal. Thus we have a current gain of 100 or even 500.

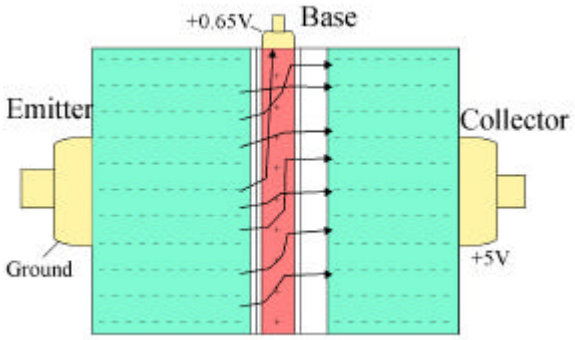


Fig 1-2: The electrons in the base of an NPN transistor are intended to flow to the base terminal but, if the base is very thin, most of them are diverted by the positive potential of the collector.

The bipolar transistor is an odd amplifier, quite non-linear and somewhat difficult to use. Consider the input terminal, the base. It is a diode (with respect to the emitter). You need to lift its voltage up to at least 0.6 Volts (at room temperature) for any current to flow. From that point on the current increases

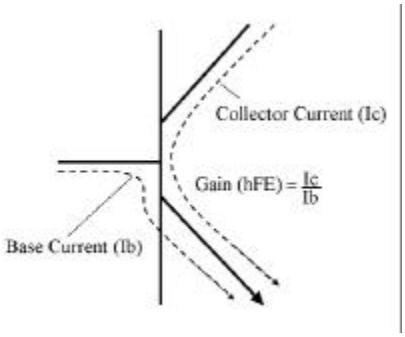


Fig. 1-3: The current flow and gain of an NPN transistor.

exponentially, both in the base and the collector. It is not a linear voltage amplifier; only the currents have a (more or less) linear relationship.

Also notice that the emitter current is always larger than that of the collector, since it contains both the collector and base current.

We have shown here an NPN transistor. If we reverse all the doping and the voltages we create a PNP transistor. It works the same way in every respect except that it is a bit handicapped: it is slower and has a lower gain; holes, now the minority carriers in the base, just don't move as well as electrons.

The point-contact transistor was a nightmare to manufacture and had very poor reliability. Also, these devices were made from germanium, which has a rather limited useful temperature range. The junction transistors were made by alloying dopant materials on either side of a flat piece of germanium or silicon. It was difficult to make the base uniformly thin and the process created considerable leakage current.

The next big step was again invented at Bell Labs: diffusion. At room temperature gases mix even if they are held perfectly still. This happens because each atom or molecule moves around randomly due to the energy it receives by temperature. The higher the temperature, the more pronounced is this movement and thus the mixing or diffusion. If the temperature is high enough (e.g. over 1000^{o}C) such gases can even diffuse into solid material, though their diffusion speed decreases enormously. Thus, for example, silicon exposed in a high-temperature furnace to n-type impurity (gas) atoms develops an n-layer at its surface with a depth as far as the impurities penetrate. This may require a temperature close to the melting point of silicon and take several hours for a penetration of just a few micro-meters, but it is far more controlled than alloying.

Moreover, you can dope repeatedly. Suppose you have a piece of silicon which has been doped n-type. If you diffuse p-type impurities into the surface, you convert a layer from n-type to p-type if there are more p-type impurities than n-type. The junction is located at the depth at which the two impurities are equal in concentration. A second diffusion of a yet higher concentration can then convert the material back to n-type again. However, you have to pay attention to the fact that subsequent exposure to high temperature causes any previous layer to diffuse further.

There are a few more dopants available too: p-type gallium (rarely used) and n-type arsenic and antimony. The latter two have the advantage that they diffuse more slowly than phosphorus or boron. For this reason they are primarily used early in the process and are thus less affected by subsequent diffusions.

When, in 1956, the three inventors of the transistor were awarded

the Nobel Prize for physics, only Walter Brattain was still at Bell Laboratories. John Bardeen had left in 1951 to become a professor at the University of Illinois and, for his research there in superconductivity, he received a second Nobel Prize in 1972.

Bill Shockley left Bell Labs in 1954. Banking on his reputation, which had risen proportionally to the acceptance of the transistor, he managed to strike a deal with the Beckman Instruments Company. A subsidiary, called the Shockley Semiconductor Laboratories, was set up in Palo Alto, California. Shockley's fame had risen to such a height that he could pick some of the best people. Within a year he had some 20 people - predominantly Ph.D.s - working for him, among them Robert Noyce, 28, Gordon Moore, 27, and Jean Hoerni, 32.

For all of these people there was a brief period of fascination after they joined. But then the true Bill Shockley appeared from behind the glitter of fame and they discovered that Shockley was, in fact, a rather erratic and unpleasant man. He would fire his employees for minor mistakes, throw tantrums over trivial problems and change directions for no apparent reasons. He incessantly tried innovative management techniques, such as posting everybody's salaries on the bulletin board.

Noyce and Moore were pushing Shockley to make silicon transistors using the diffusion approach. Shockley wasn't interested; his hope was for his laboratory to come up with an entirely new device, a device which would represent as large a step over the transistor as the transistor had been over the vacuum tube.

Now totally dissatisfied, the crew talked to Arnold Beckman, the president of the parent company, and informed him of the impossible situation. Beckman promised to hire a business-minded individual who could act as buffer between Shockley and his staff. But the solution didn't work, Shockley refused to let go of the day-to-day decision-making. Out of patience, eight staff members reached a deal with the Fairchild Camera and Instrument Company and, in October 1957, the group departed.

The new company, called Fairchild Semiconductor, was at first an independent operation, with Fairchild Camera and Instrument holding an option for a buy-out. The product they began to develop was the one they had proposed to Shockley. The detailed structure of this device, called the Mesa transistor, had been tried in germanium before, but not in silicon. It required two diffusions, both into the same side of a silicon wafer. The first diffusion was p-type, the second n-type, and the difference in depth between the two layers created the base region which, for the first time, could be made with a high degree of accuracy. The top surface of the transistor was then masked with wax and the exposed silicon etched away, giving the

remaining piece a mesa-like shape.

Because of its superior performance, sales of the Mesa transistor took off almost immediately, reaching $ 7 million in 1959. But there were also problems. The most serious one concerned the reliability of the Mesa transistor. The etched silicon chip was soldered onto the bottom of a small metal case, leads were attached to the top regions and then the case was welded shut. Tiny metal particles, ejected during the welding process, floated around inside the case and kept on shorting out the exposed p-n junctions.

Silicon rapidly grows a thin oxide layer when it is exposed to air. This is better known as glass (silicon-dioxide) and its growth can be enhanced by moisture at high temperature. Some of the dopant gases used in diffusion (such as gallium) can penetrate this oxide layer, while others are stopped by it. There was, therefore, a possibility that the oxide layer could be used as a mask. If the oxide were to be etched off in some places but not in others and suitable dopant gases used, diffusion would take place only in the areas without oxide. But a study done at Bell Laboratories came to the conclusion that an oxide layer exposed to a diffusion is left contaminated and must subsequently be replaced by a freshly grown one.

This bothered Hoerni. He didn't see any reason why the oxide layer could not be used as a diffusion mask for both diffusions -- provided he would use dopant gases which were stopped by the oxide - and why the oxide should subsequently be regarded as contaminated. So he tried it -- as an unofficial side project -- and out of the trial came an advance ranking in importance second only to the transistor itself: the *planar* process.

In preparation for the first diffusion Hoerni spread a photosensitive and etch-resistant coating (photoresist) over the top of the oxide and exposed it through a photographic plate (mask) carrying the patterns of the base regions, using the photographic techniques already developed for "printed" circuits. The subsequent etching then only removed the oxide in the regions where p-type impurities were to be diffused. After the diffusion he closed these oxide "windows" again by placing the wafer in high-temperature moisture and then repeated the steps for the second (emitter) diffusion. In a third masking step windows could then be etched in the oxide to make contact to the two diffused layers. He then evaporated aluminum onto the top surface of the wafer and patterned it with the same photographic techniques. The wafer could then be scribed (like glass) and broken into individual transistor chips.

The planar process had a whole series of advantages. Of most immediate importance was the fact that the junction was automatically protected by the oxide, one of the best insulators known. No longer could

the metal particles from the welding of the case short it out. Secondly, photographic methods could be used to delineate not just one but hundreds of transistors simultaneously. Thus individual, delicate masking of each transistor was no longer required, giving the planar transistor a huge potential for reduced cost. Noyce, who was by now the general manager, saw the advantage of the planar process and quietly moved it into production.

There was another advantage to the planar transistor: once the dopant enters the silicon it diffuses in all directions, including sideways. The P-N junction, therefore, ends up underneath the oxide, never exposed to either human handling or the contamination of air. For this reason the planar junction is the cleanest (and most stable) junction ever produced. Fairchild's customers who, in early 1959, didn't know that their transistors were now being manufactured by an entirely new process, were surprised to find leakage currents one thousand times smaller than those of previous shipments.

While Fairchild flourished, Shockley Transistor went downhill. It was sold twice, then closed in 1969. Shockley became interested in sociology and announced a theory called "dysgenics", which proposed that poor people were doomed to have low IQs. By the time he died in 1989 his reputation was ruined.

The Integrated Circuit

In July 1958 Jack Kilby of Texas Instruments conceived that a block of germanium or silicon could be host to not only transistors and diodes, but resistors and (junction) capacitors as well. This appeared to be enough of a variety to make a small circuit, all of it in the same block of silicon.

The idea was good, but his approach cumbersome. To insulate the various components from each other Kilby etched the silicon, in some areas all the way through. To connect them together he used gold wires. The circuit was very small to be sure, but it was a production nightmare. Each tiny block of silicon had to be made individually, including the patterning, etching and wiring. When TI's attorneys prepared a patent application they looked in horror at the Rube Goldberg-like drawings and had Kilby put in some words saying that interconnection could also be made by laying down a layer of gold. How this could be done over this three-dimensional landscape he didn't say.

While Kilby was working on his circuits in Texas, a similar but far more elegant idea occurred to Robert Noyce in California. Noyce's

motivation was primarily cost, not size. He realized that it didn't make sense to fabricate precisely arranged transistors on a wafer, cut them apart, place them in a housing and arrange them again in on a circuit board; if the additional components on the circuit board could be placed on the wafer, a considerable number of manufacturing steps could be saved. Noyce had no problem visualizing capacitors and resistors made in silicon, he was constantly dealing with these (unwanted) effects. What was needed, though, was an inexpensive way to connect all these components on the wafer. The idea of using wires had no chance in Noyce's mind, it would have simply been too expensive. But he saw that, in the planar process, this problem was already solved: the aluminum layer used to connect the transistors and the wires could also be used between the components.

In1959 Noyce entered his idea into his notebook and filed for a patent application. Kilby's and Noyce's patent applications were clearly in interference and a bitter battle between the two companies started in the courts. Texas Instruments won because Kilby application mentioned a thin film of gold, thus seemingly anticipating Noyce. Fairchild appealed.

While the two patents were fought over in the courts, neither TI nor Fairchild could collect any royalties for integrated circuit, which were already showing explosive growth. So the two companies came to an agreement, declaring Kilby and Noyce co-inventors of the integrated circuit. Shortly after this the appeals court handed down its decision: Noyce, not Kilby, was declared the inventor of the IC.

It could not have been otherwise. Even today every single IC is made exactly as Noyce described it, while Kilby's approach has long been abandoned. But the most important contributor to the invention of the IC was clearly Jean Hoerni with his planar process, for which he has never been adequately recognized. The planar process rates as one of the great inventions of the 20th century.

Robert Noyce died in 1990 at age 62. In 2000 Jack Kilby won the Nobel Prize for the invention of the integrated circuit

Let's take a closer look at a basic processing step in the Planar process. First, you need a mask, a piece of flat glass, with an opaque pattern on it. The pattern has been generated optically or, more likely, with an electron beam.

The silicon wafer is first oxidized, i.e. a thin SiO_2 layer is grown, for example by exposing the wafer to steam in a furnace. Instead of oxide, nitride or a combination of oxide and nitride is also used. On top of the oxide a thin layer of "photoresist" is spread, a light-sensitive emulsion similar to that on a photograph.

Light is then projected through the mask onto the wafer. The higher the frequency of the light, the greater the detail, so ultra-violet light or even x-rays are used.

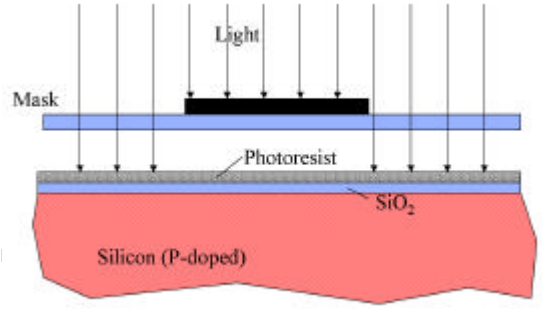


Fig. 1-4: The first step: A light-sensitive and etch-resistant layer (photoresist) is spread on the wafer and exposed to light through the mask.

The photoresist is then developed and the portions not exposed to light are washed off. (There are both positive and negative photoresists; you have the choice of removing the areas which are either exposed or not exposed to light).

Fig. 1-5: The photoresist is developed like a photograph and the wafer is ready for etching.

Next the entire wafer is immersed in an acid which removes the oxide in the areas where it is not protected by the photoresist. In more modern processes a plasma is used; acid etches not only downward but also slightly sideways underneath the photoresist, while plasma etches downward only.

Fig. 1-6: The oxide is etched away and the photoresist is removed.

The wafer is then placed into a furnace (a quartz tube heated to greater than 1000°C). A gas carrying the desired dopant (in this case boron, arsenic or antimony) swirls around the wafer and slowly diffuses into the surface.

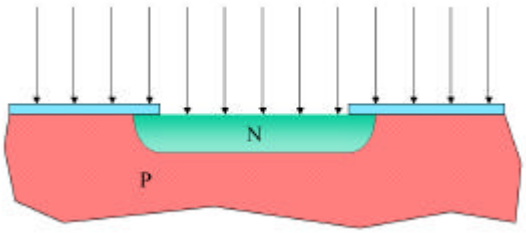


Fig. 1-7: A gas containing N-type dopants (boron, arsenic or antimony) diffuses slowly into the surface of the wafer at high temperature.

Note two important facts here: 1. There is a crowding of dopants near the surface of the silicon. With time they will diffuse deeper into the silicon, but there will always be more dopants near the surface. Thus any diffused region has a marked gradient. 2. Dopants not only diffuse downward, but also side-ways. (Since supply is more

limited at the very edge, the side-ways diffusion extends to only about half the distance of the downward one). This places the junction (where n = p) underneath the oxide and is thus never exposed to the (dirty) environment.

After diffusion the exposed silicon surface is covered again by an oxide layer so that the wafer is ready for the next masking step, which could be another diffusion or the etching of contact holes.

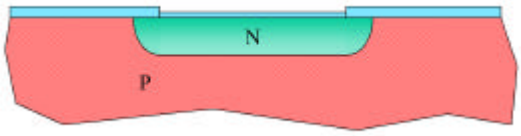


Fig. 1-8: After the diffusion the oxide is re-grown, ready for the next masking step.

There is an important feature here, which should not go unnoticed. SiO_2 is glass, which is transparent to light. The light is reflected at the bottom of the oxide by the silicon and interference patterns are created, i.e. the sum of direct and the reflected light eliminates some frequencies. Thus the color of the oxide layer depends on its thickness. This not only makes for beautiful photographs but, more importantly, it allows subsequent masks to be precisely aligned with previous ones.

Here then is one form of an NPN transistor made with the planar process. The substrate (the starting wafer) is doped p-type as the silicon is grown. There are three diffusions in succession, the first being rather deep. After the diffusions, contact holes are made (with the same basic photoresist process), aluminum is deposited over the entire wafer, patterned (another photoresist step) and etched away where it is not wanted.

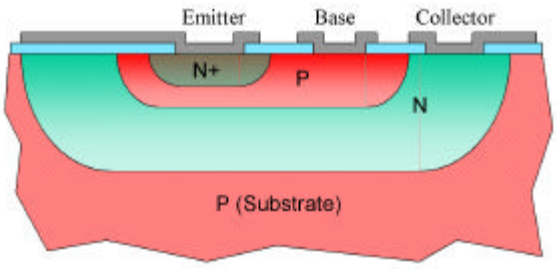


Fig. 1-9: A simple planar NPN transistor.

Alas, this transistor has a rather significant shortcoming: high collector resistance. The current has to flow through the region between the base and the substrate. That is the far end of the collector diffusion, the end which has the fewest dopant atoms and therefore the highest resistance.

Since the invention of the planar process a few more ways of fabricating have been added:

Epitaxy. If you strip a silicon wafer of its oxide and put it into a furnace which is filled with gas containing not only a dopant but also silicon, you can grow a doped single-crystal layer. As the atoms carried by the gas deposit themselves on the surface of the wafer, they will align

themselves according to the existing crystal structure.

You can also precede this by diffusing regions into the original wafer, so that you will have areas of high concentration underneath the **epitaxial layer**. Even though these regions are buried, it is still possible to align subsequent diffusions to them. When a diffused area is re-oxidized, a small amount of silicon is consumed (the Si in SiO_2), thus creating a small depression in the surface. The edges of these depressions are visible at the top surface of the epitaxial layer, though the image tends to be blurry and is shifted (in most processes) along the crystal axis (around 45^o).

Ion Implantation. You can literally shoot dopant atoms into silicon by electrically charging (ionizing) them and then accelerating them with a high voltage (several hundred thousand volts). The treatment is somewhat brutal, the newly arrived atoms don't end up neatly aligned in the crystal structure and an annealing heat cycle is necessary to let the atoms align themselves into a crystal structure.

The number of dopant atoms introduced is generally more accurate in ion implantation than in diffusion. Also you can aim implantation for a certain depth (but not very deep). In the subsequent heat cycle (and during subsequent diffusions) the dopant atoms will diffuse and thus widen the layer. The maximum concentration, however, is then not at the surface, but at a chosen depth.

We now have arrived at a modern NPN transistor as made in a bipolar (or BICMOS) process. Before growing the epitaxial layer, a heavily doped (thus N+) **buried layer** is diffused (or ion implanted) into the p-type substrate. During epitaxy it diffuses somewhat, both into the substrate and the new epitaxial layer.

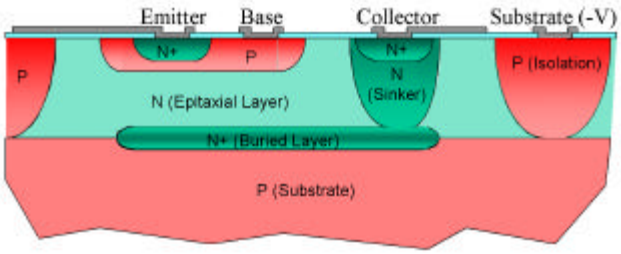


Fig. 1-10: A much improved planar, integrated NPN transistor. The buried layer and sinker lowers the collector resistance.

The next diffusion is the **isolation**. It is deep (and, therefore, also wide); it has to connect up with the substrate, so that the entire n-type collector region is surrounded by p-type regions. A second n-type diffusion connects up with the buried layer (and the emitter N+ diffusion is used on top of it simply because it's available at no cost). Now the collector current has a (fairly) low-resistance path.

This transistor is isolated from its neighbors (and other components) as long as the substrate is held at the most negative voltage in the circuit (**junction isolation**). In this way the collector-substrate junction is always reverse-biased and only leakage current (pico-amperes) flows.

There are some flaws and limitations in the performance of this or any other bipolar transistor:

Early Effect, named after Jim Early (then at Bell Labs, later at Fairchild), who explained it first. Ideally the collector current should be equal to the base current multiplied by a constant gain (hFE or beta). But, as we have seen above, each p-n junction has two depletion layers. For the collector-base junction, one depletion layer extends into the collector, the other into the base. The base is almost always more heavily doped than the collector, so its depletion layer is fairly shallow. However, the base is also very thin, so even a shallow depletion layer takes up a significant portion of the base depth. As the collector voltage increases, the depletion layers widen. In the collector region this has little effect (as long as it doesn't hit the other side of the collector), but in the base region it *narrows* the base-width. Since the gain of a bipolar transistor is very much dependent on the base-width, the gain simply increases as the effective base-width decreases.

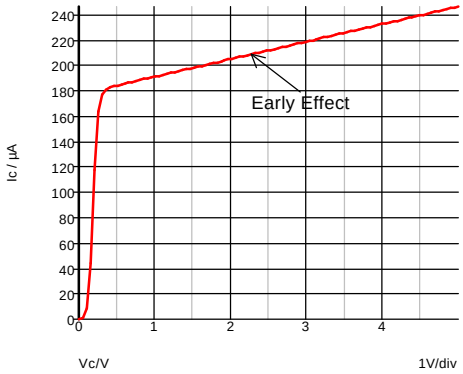


Fig. 1-11: Even with a constant base current the collector current increases with the collector-emitter voltage because the depletion layer narrows the base-width.

If you draw a straight line, extending the slope (from 0.4 to 5 Volts) into the negative quadrant and let it intersect with the zero-current line, you get the **Early Voltage**. In this case, for a 5-Volt process, the Early voltage is -15 Volts (but is generally expressed as 15V). Depending on the chosen base-width, it can be less than that and the slope correspondingly steeper.

Gain versus Current. For any bipolar transistor the current gain falls off both at low and high current.

First, the low end. There is always a leakage current across any junction; for a perfectly clean surface this is the diffusion current. In the base-emitter junction this leakage current takes away a portion of the supplied base current. In our graph here the current shunted by leakage at

the low end (10nA Ie, or about 50pA Ib) amounts to 33% of Ib, i.e. the gain has dropped by one third.

If you extend this plot to much lower current, you will see the gain rise to almost infinity. This is nothing more than the effect of the collector-base leakage current.

At the high end two effects take place simultaneously: 1. The number of electrons present in the base simply becomes so large that they are no longer the minority carriers and the whole effect comes to a halt. 2. The base current must flow from the contact to the flat area between the emitter and collector. At low current this is no problem, the resistance in the base is sufficiently small. But as the collector current increases (and with it the base current), the resistance in this flat region of the base causes a significant voltage drop, and the far end gets less current. Eventually, as the current is increased even more, only the edge of the emitter on the side of the base contact is active. Thus the high-current capability of a bipolar transistor is determined not by the emitter area, but by the **active emitter length**, i.e. emitter periphery to which the base can supply current through low resistance. A good starting point for the maximum current (at which the gain drops to 50%) is 1.5mA per um of active emitter length, but this value varies from process to process.

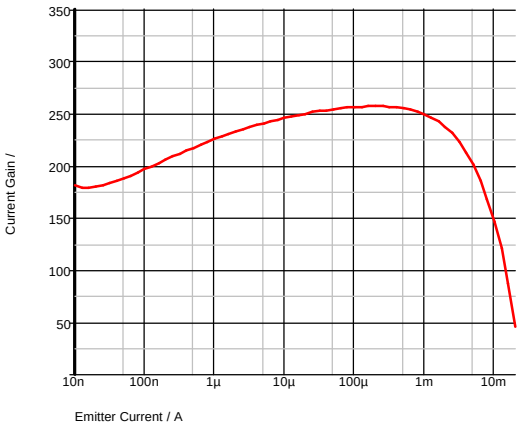


Fig. 1-12: The current gain (hFE) of a bipolar transistor drops off both at low and high currents.

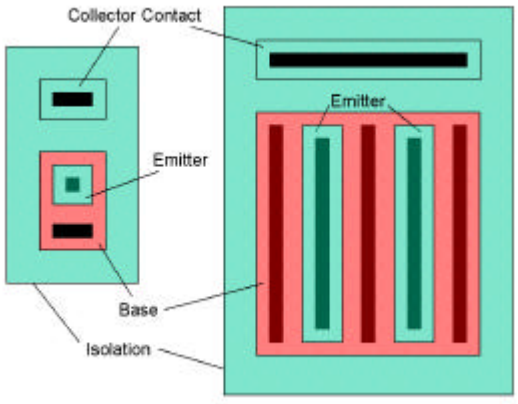


Fig. 1-13: Minimum-geometry NPN transistor on the left and higher-current design on the right.

To increase the current capability of a bipolar transistor you can

place base contacts on both sides of the emitter and lengthen the emitter. Shown here on the left is the top view of a minimum-geometry transistor and on the right a version for higher current.

To make the life of a designer easier, the isolation pattern is usually drawn as a rectangle and then inverted when making the mask, i.e. the isolation diffusion is actually *between* devices, not in the device area.

Many processes require that all contacts be the same size, in which case the contact rectangles must be broken up into small, identical (and properly spaced) squares.

Be aware, that transistors of different sizes (as drawn here) do not match well. At low current a large emitter area produces a higher gain than a small one, because the minority carriers have a higher chance to be captured by the collector. If you want to produce a precise ratio, use only one emitter size and identical base contacts. The emitters can be in a common base area and the collector size is of no consequence except for collector resistance (or saturation voltage).

Substrate Current. There is only leakage current across the collector-substrate junction, unless the transistor saturates.

Assume the collector is connected through a resistor to the positive supply voltage and the base is driven so hard that the collector voltage drops to near the potential of the emitter (termed saturation).

There are now two diodes in parallel and the base current has two paths; the new one forms a PNP transistor with the NPN base becoming the emitter, the NPN collector the base and the substrate the collector. Since the NPN collector is much larger than its emitter, some (or all) of the base current flows to the substrate.

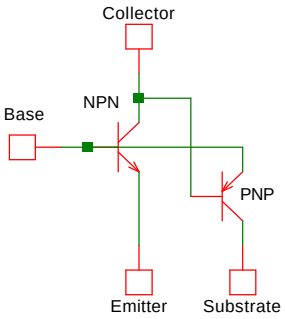


Fig. 1-14: When an NPN transistor saturates a stray PNP device leaks current to the substrate.

There is little danger in this, except when you drive the base very hard, trying to get the lowest possible collector voltage, or if you have many saturating NPN transistors. The path in the substrate from a transistor to the -V connection has some resistance. If the substrate current is so large that the voltage drop across this resistance can forward-bias some substrate-collector junction on the way, you may get some really bad effects, including latch-up.

Maximum Voltage. To get a high operating voltage requires high resistivity - low doping concentration. But there is a price to be paid: the depletion regions become wide.

Let's use the integrated NPN transistor as an example. There are two depletion regions, one extending into the epitaxial layer from the base (downward and side-ways), the other into the epitaxial layer from the isolation. To make sure the first one does not reach the substrate (and thus cause premature breakdown or **punch-through**), the epitaxial layer must be deep - which means that the isolation diffusion must be deep, and thus wide.

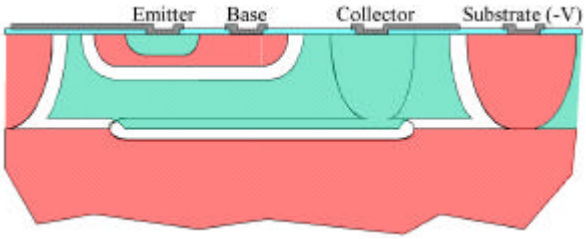


Fig. 1-15: At higher operating voltage the depletion regions around the NPN transistor become larger.

Look at the left side of the transistor. The spacing between the isolation (as drawn) and the base must accommodate the following:

the side-ways diffusion of the isolation,
the isolation-collector depletion region,
a safety margin for possible misalignment,
the collector-base depletion region and
the side-ways diffusion of the base.

In addition there is also a high-voltage depletion layer each between the base and the sinker and the sinker and the isolation, as well as a deeper (and thus wider) sinker. All this adds up to a painfully large area.

The increase in area can be curbed somewhat by two measures: 1. Use an additional diffusion for the isolation by creating a P+ region directly underneath the normal one before growing the epitaxial. The two halves will then diffuse toward each other (**up-down diffusion**) and meet in the middle, thus requiring only half the depth and width; 2. Add more processing steps, creating both low-voltage and high-voltage devices on the same wafer.

The Miller Capacitance. As we have seen above, the bipolar transistor is a very non-linear (exponential) voltage amplifier and cannot thus normally be used as one. But it *has* a voltage gain, and a high one at that (several hundred is not uncommon).

There is an unavoidable junction capacitance between the collector and the base. If you feed a current with an ac signal into the base, the voltage change on the collector will be much larger than that on the base. Thus, looking into the base, the junction capacitance appears *multiplied* by the voltage gain (the Miller effect). Instead of a tiny fraction of a pico-Farad you have to deal with 10 or even 100pF. If the base is fed from a high impedance (e.g. a current source), the frequency response is then

nowhere near the advertised f_t (cutoff frequency).

This Miller effect can be reduced by circuit design techniques (e.g. a **Cascode Stage**), but even so most circuits cannot operate much above say 1/20 of f_t. (f_t is the frequency at which the current gain drops to 1).

On the other hand, there is also a benefit. In feedback amplifiers you almost always need a **compensation capacitor** (more of this later). Using the Miller effect you can get away with a 5pF capacitor, which appears to be as large as 1nF, a value which would be much too large to be integrated.

The Case of the Lateral PNP Transistor

It is the world's worst transistor, you couldn't sell it as a discrete component: low cutoff frequency, very limited current range and an inferior noise figure. But no self-respecting analog IC designer would want to be without it. The reason: In either a CMOS or bipolar process no additional diffusions are required.

The emitter and collector are formed by the base diffusion (in a bipolar process) or the p-channel source/drain diffusion (in a CMOS process). The current thus flows radially (or laterally) along the surface from the emitter to the collector.

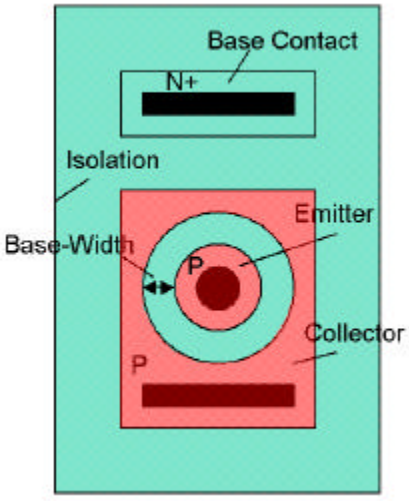


Fig. 1-16: The lateral PNP transistor.

The doping levels are all wrong. For optimum performance you would want the emitter to have a very high concentration, the base somewhat lower and the collector quite low (to accommodate the higher collector voltage). Here the emitter and collector doping levels are equal, and the base is much higher. Thus, to allow space for the depletion regions, the base width (the distance between the collector and emitter, minus the side-ways diffusions) needs to be quite large. Hence the slow speed (it takes time for the carriers to travel across the base). Figure on an f_t in the neighborhood of 30MHz with an operating voltage of 15 Volts; at lower voltages the base-width can be narrower which increases f_t but also makes the Early effect more pronounced.

Despite of all of this, with good surface control you can get a gain in excess of 100. But the current range is limited, rarely exceeding 100uA for a minimum geometry device.

And there is somewhat of a problem with substrate current. There is a competing PNP transistor, using the same emitter and base, but with the substrate (and the isolation diffusion) as the collector. In normal operation a current about half the magnitude of the base current flows from the emitter to the substrate terminal. When the lateral PNP transistor saturates, the substrate current becomes almost equal to the collector current. If you don't have a buried layer, it gets quite a bit worse.

One advantage of the lateral PNP transistor: the collector can be split into two (or more) sections. The emitter current, flowing radially outward is collected by the segments according to their length at the inside. There is a small loss in gain because of the gaps, but the matching between the two collector currents is excellent.

In a CMOS process emitter and collector are usually formed by the p-type diffusion of a p-channel MOS transistor. The intervening space (the base-width) is the same as a p-channel gate, with poly-silicon on top. Connect the poly region to the PNP emitter; it will act as a static shield and have a (slight) beneficial effect.

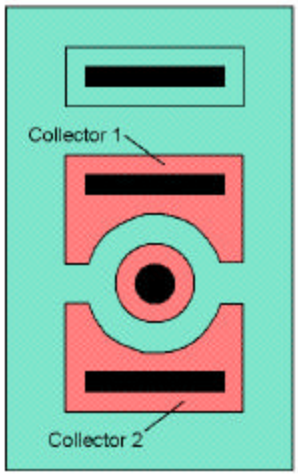


Fig. 1-17: A split-collector lateral PNP transistor.

CMOS Transistors

It took almost 20 years after the invention of the bipolar transistor for MOS to make its appearance. Shockley (and many others) had thought of this device first, it was (or should have been) much more simple: put a plate close to the surface of silicon, connect it to a voltage and move the carriers inside the silicon electro-statically.

The problem was the surface of silicon. Here the silicon atoms are no longer neatly tied up with each other by sharing the outermost electrons. They face an entirely different material, SiO2 (or worse, some covering with unknown impurities mixed in). This material doesn't even have a crystal structure, it is amorphous.

In 1964 a startup, General Microelectronics, felt it had licked the problem with CMOS and brought out the first digital MOS integrated circuit. It was one of the worst products ever to hit the market: a large portion stopped working within days. The reason: there were elements with the silicon-dioxide (chiefly sodium) that carried an electric charge and could move. One day the MOS transistor was perfectly functional, the next day it was permanently turned on.

It took another few years to gain an understanding of MOS surface physics and make stable MOS transistors. Today the silicon surface is so well understood that we can deliberately place a charge into the oxide layer that stays there for years, probably even centuries. It is now the dominant integrated device, being much smaller than the bipolar transistor. (The number of MOS transistors produced every year has long surpassed the number of ants in the world. At the time of writing this book, semiconductor manufacturers produced some 500 million transistors for every person in the world per year).

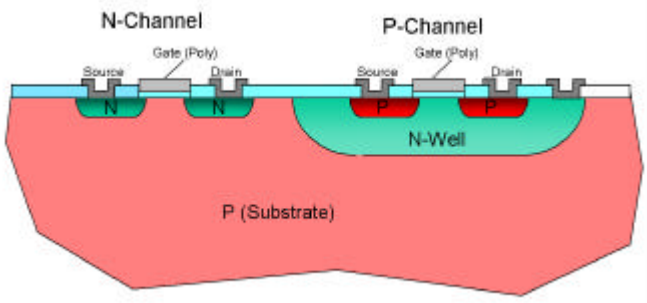


Fig. 1-18: Cross-section of an N-well CMOS process.

The figure shows a cross-section of the most often used (n-well) process. There are many variations and refinements; this is only the basic one.

In the gate area the insulating layer (SiO2 or nitride, or a combination) is thinned down and silicon is grown on top of it. Since the insulator is amorphous, the grown silicon is not single-crystal, it consists of many small crystals which do not fit together very well (thus it is called poly-crystalline silicon or simply **poly**).

Next the source and drain regions are implanted, using a mask. The inside edges are masked by the gate, so they align perfectly to the gate (i.e. they are **self-aligning**). The device is also **self-insulating**: as long as the source and drain are at or above the substrate potential (usually ground), the junctions to the substrate are reverse-biased and no bulky isolation diffusion is necessary.

For the p-channel transistor the polarities for the source and drain implants are reversed and these regions are placed inside an n-type diffusion. In most applications one such n-well hosts many p-channel

transistors and is simply connected to the positive supply voltage; in this way the devices are insulated from each other as long as each source and drain is at or below the positive supply.

In both the n-channel and p-channel transistors, sources and drains are identical, i.e. you can arbitrarily call one the source and the other the drain. Or one region can do double-duty, being the drain for one transistor and the source for the next one, connected in series.

The p-channel transistor is always at a disadvantage, because holes are more difficult to move than electrons. Thus it will have a lower gain than an n-channel device (for the same gate oxide thickness) and be somewhat slower. (MOS transistors, by the way, are called **unipolar** devices, because they employ only one type of carrier, as opposed by the bipolar transistor, in which both electrons and holes are important for the operation).

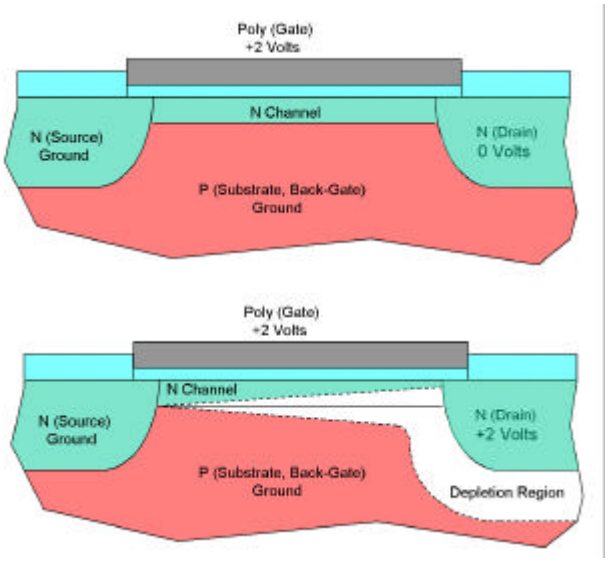


Fig. 1-19: As the drain voltage is increased, a depletion region pinches off the channel.

Now let's look at an (n-channel) MOS transistor in more detail. The basic idea is to create a region (a channel) between source and drain which has the same polarity (n-type), so that there is direct conduction between the two. This is done with a positive voltage at the gate which pushes holes away from the surface and the device is called an **enhancement-mode** transistor (there are also **depletion-mode** devices in which a channel is implanted or diffused and then cut off with a negative gate voltage).

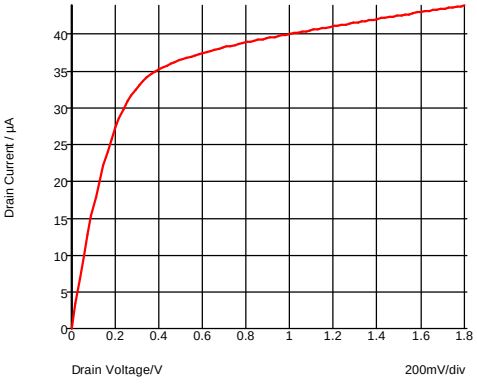


Fig. 1-20: Drain current vs. drain voltage with the gate voltage held constant.

This is true only at zero or very low drain voltage. As the drain voltage is increased, a depletion

region forms around it. Since there is now a voltage drop along the channel, with the drain side at a higher voltage than the source, the depletion region along the channel gradually increases toward the drain, cutting more and more into the channel. Thus the resistance of the channel increases.

The initial slope of the drain voltage / drain current curve is the resistance of the channel without any depletion layers. The final slope at the highest drain voltage represents its resistance with the depletion layer almost pinching off the channel. It is an unfortunate fact that this region is called the "saturation region", which clashes badly with the earlier definition for the bipolar transistor.

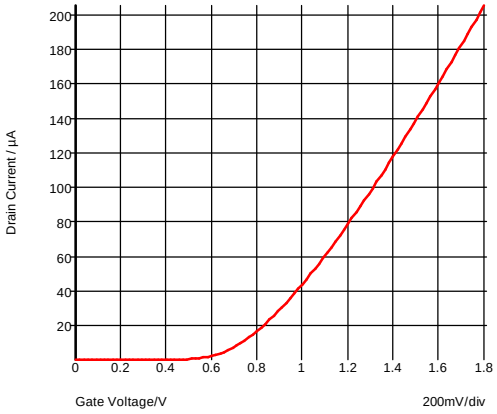


Fig. 1-21: Drain current vs. gate voltage with the drain voltage held constant.

Above a certain gate potential, which has to be exceeded to attract any carriers to the surface (the **threshold voltage**) an MOS transistor is basically a square-law device: doubling the gate voltage results in four times the drain current. The measure of gain is the **transconductance**, drain current divided by gate voltage. So again, like the bipolar transistor, this is a non-linear device:

$$Id = k\frac{W}{L}\left(Vgs - V_T\right)^2$$

where Id = drain current
k = transconductance
W = channel width
L = channel length
Vgs = gate-to-source voltage
V_T = threshold voltage
or Vgs - V_T = gate voltage above the threshold

The region below the channel also influences the gain. It forms a **back-gate**. In an n-well n-channel transistor this is the substrate, common to all devices. You have no choice but to connect it to the lowest negative

voltage. But there is a choice for the p-channel transistor. If you place all the p-channel transistors in a common n-well, you get the smallest total area and therefore the lowest cost. But if the source of such a transistor is operated *below* the positive supply, the back-gate (the n-well) pinches off the channel further and you get a reduced gain (by perhaps 30%). You can avoid this by placing this transistor in its own n-well.

The Substrate PNP Transistor

In either a bipolar or CMOS process there exist layers which can form a PNP transistor with the substrate as the collector. Since the collector is permanently connected to the most negative supply voltage, such a device has limited use. In a bipolar process a lateral PNP transistor has greater flexibility and better performance and is thus almost always preferred.

In a CMOS process the same is true, but because of historical reasons or limited information the substrate transistor is still present. The p-type implant for the p-channel transistor forms the emitter, the n-well the base and the substrate the collector. The n-well has a large depth, thus the PNP base-width is large and the gain rather small (e.g. 10).

Diodes

There are several p-n junctions in an integrated circuit, each and every one a diode. But few of them can actually be used by themselves without unpleasant side-effects.

Take a simple bipolar process. There are three types of junctions: emitter/base, base/collector and collector/substrate (all referring to the NPN transistor). The last one is hardly ever useful because the substrate is permanently connected to the most negative supply voltage. The base/collector diode is, as we have seen, part of a substrate PNP transistor with a gain; a current perhaps ten times the magnitude of the diode current will flow to the substrate.

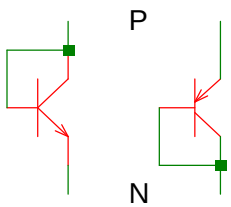


Fig. 1-22: Properly connected diodes in a bipolar process.

The emitter/base junction makes a good diode, but it has a low breakdown voltage (about 6 Volts) and the base has a fairly high resistance. You

could connect the surrounding collector to the most negative supply voltage and thus keep it always reverse-biased. But a much better diode results if you short collector and base together, creating a **diode-connected transistor**. The transistor is active, it has gain. Only a small fraction of the current flows through the base, which divides the base resistance by the current gain. This connection in fact gives you an almost ideal diode over about five decades of current.

If the emitter/base breakdown voltage is too low, consider a diode-connected lateral PNP transistor. This devices has the full operating voltage of the process, but it is limited in current (see above).

In a CMOS process the restrictions are even more severe. The only free-floating junction is between the p-channel source/drain and the n-well. But, as we have seen, these are also part of the substrate PNP transistor. Were you to run a current through this junction, a current of about ten times its magnitude would flow to the substrate.

The term "diode-connected" is often used for an MOS transistor with its gate and drain connected together. Don't be misled by this term: there is no diode as in "junction" diode.

Zener Diodes

In a bipolar process the base-emitter diode almost always has a low breakdown voltage (perhaps 6 Volts) with a fairly low temperature coefficient, which makes it useful as a reference voltage.

But exercise care with this device: the same junction is also used as a fusible device.

At low current (e.g. less than 100uA for a minimum-geometry device) the Zener diode behaves well. As you increase the current the region between the emitter contact and the edge of the emitter diffusion lights up faintly (a plasma, which you can observe under a microscope, with all lights turned off). At some high current level a thin aluminum strip is formed abruptly underneath the oxide, which converts the Zener diode into a short-circuit. This effect is used for trimming and carries the earthy name Zener-zapping.

Such a Zener diode is also somewhat noisy. For lower noise (and better accuracy) use a bandgap reference.

Moving an n-channel and p-channel source/drain diffusion in a CMOS process close together so that they intersect can also result in a useful low breakdown voltage, but data for such a device are rarely available from the wafer-fab.

There are also **buried Zener diodes**, devices with a special diffusion below the surface of the wafer. Such devices have lower noise, but the addition to the process tends to be costly.

Resistors

Every free-floating layer in an integrated circuit can, when properly patterned, become a resistor. But for all of them this is only a secondary duty; their intended application is in a transistor, which is the hardest device to make. It shouldn't come as a surprise then that their values have a higher variation and greater temperature coefficient and their range is more restricted than that of even the least expensive discrete resistor.

Discrete resistors can be tested and adjusted during manufacturing. In ICs the manufacturing is done while the silicon is red-hot, at which temperature it is no longer a semiconductor; you have to wait until it cools down to measure any parameter.

What saves the integrated resistor is its natural ability to match well. Whatever error may have occurred in making one applies to any other on the same wafer. They may both be as much as 25% high in value, but both will be high by (almost) exactly the same amount.

The resistance of any material is given by

$$R = \frac{rho \cdot l}{A}$$

where rho (ρ) = resistivity in Ohm-cm
l = length
A = area (cross-section)

If we make a square, i.e. w = l, then we get a measure of resistance which is independent of size, the **sheet resistance**, in Ohms per square (or Ohms/□).

Note the term is sheet resistance, not sheet resistivity. A square in a layer with a sheet resistance of 100 Ohms per square always measures 100 Ohms from one side to the other no matter how large the square.

In a bipolar process the layer most often used for resistors is the (NPN) base (about 200 Ohms/□). The emitter layer is more heavily doped and thus has a lower sheet resistance (as low as 5 Ohms/□).

In a CMOS process you have a wider choice: the n+ and p+ diffusions (implants) for the drains and sources, the n-well and usually two different poly layers. Of these the p+ diffusion (about 150 Ohms/□) and one of the poly layers (around 50 Ohms/□) are generally best suited.

Sheet resistances depend greatly on the process; you should use the

values given here only as a starting point and get the actual data (including temperature coefficients and tolerances) from the wafer fab.

Diffused resistors must be placed in an island of opposite doping and this island must be connected to a bias voltage so that the junction is reverse-biased. For example, a (p-type) base resistor must be in an n-type (epi) island. This island (sometimes called the "tub") can contain just one resistor or all of them, but its voltage must be at a level equal to or greater than the largest voltage on any resistor. In this case the easiest and safest connection is to +V.

Diffused resistors (and to a lesser degree, poly resistors) have a **voltage coefficient**. The biased surrounding layer pushes a depletion region into the resistor, reducing its cross-section. As the difference in voltage between the resistor and the surrounding layer becomes larger, the depletion region widens, the cross-section becomes smaller and the resistance increases. This effect is especially pronounced in lightly doped layers: the n-well in CMOS and **implanted resistors** in a bipolar process. (The latter uses an additional implant to create a high sheet resistance).

This voltage dependence is especially critical if you have two (or more) resistors which need to match but are at different DC levels. You can place each resistor in a separate island, biased at the positive end of its resistor. Or you can simply accept the change caused by the depletion layer and adjust the ratio. For this, however, you need a model for the resistor which includes its voltage dependence. (in a 200 Ohms/☐ base layer, for example, the change in resistance is about 1% for a 5V bias difference).

There is also a (distributed) capacitance associated with an integrated resistor, low for poly, higher (and voltage dependent) for diffused ones. If you make a high-value (i.e. very long) resistor, this stray capacitance can seriously cut frequency response. Also, if there is noise on the supply which biases the surrounding region for diffused resistors, it will be capacitively coupled into the resistor. Again, a good model is required to show these effects in a simulation.

Two correction factors have to be used when designing a resistor. The first concerns the width of the resistor. In diffused (or implanted) resistors there is always a sideways diffusion, which makes the actual resistor wider than drawn. The effect of the side-ways diffusion is dependent on the width of the resistor.

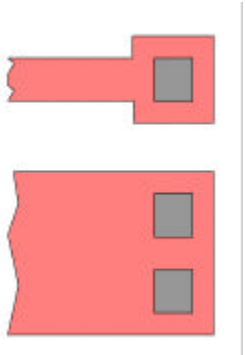


Fig. 1-23:
Resistor contacts.

The second correction factor recognizes the **end-effect**. If the resistor has minimum width, you

will need to enlarge both ends to place a contact inside. You will then need to estimate the resistance of this additional area and of the contact itself (totaling perhaps 0.4 squares from the end of the narrow part).

If you draw a wide resistor, the contacts can be fitted inside the resistor, but they will not cover the entire width, even if converted to one long contact. There is, therefore a small additional resistance (about 0.2 squares from the inside edge of the contacts).

The **matching** of resistors depends entirely on the width. Sub-micron processes are not developed to get good matching, just maximum speed. You will find that minimum-dimension devices (all devices, not just resistors) match very poorly. When greatly magnified under a microscope all edges appear somewhat ragged. The width of a resistor, for example, fluctuates considerably. It is only when you make a relatively large device that these fluctuations become insignificant and thus devices match well. Figure on using something like *ten times* the minimum width for matching of 0.5% or better.

Because of the end-effect you cannot expect resistors of different lengths to match well. For optimum matching use only *identical* resistors. It also helps divide resistors into identical sections and intermingle them with other resistors (in the same identical sections) which are intended to match.

One more thing about IC resistors: the **Seebeck** effect. Discovered in 1821 by Thomas Seebeck (and used by Ohm four years later for his measurements of resistance), it is the thermocouple effect: metallic interfaces at the ends of a wire produce a voltage if the ends are at different temperatures. For the contacts of a diffused or poly resistors this voltage is between 0.2mV/^{o}C and 1.4mV/^{o}C, depending on the doping level and composition of the metal. This is a danger if thermal gradients are present, e.g. with a power transistor on the chip. To avoid it, lay the resistor out so that beginning and end are close together.

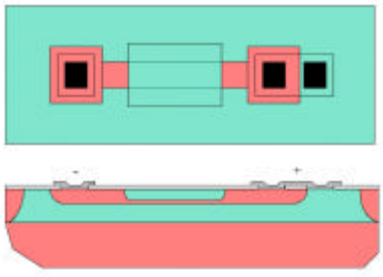


Fig. 1-24: Top view and cross-section of base-pinch resistor.

Pinch resistors (or pinched resistors) are sometimes used in bipolar processes to get a high resistance without wasting a lot of area. The **base-pinch resistor** is simply a base resistor with the emitter diffusion placed over part of it. This reduces the effective cross-section (only the deepest part of the base diffusion is left, which has also the highest resistance). The device needs to be in its own epi island, with the epi (and emitter

diffusion) connected to the positive terminal. A base-pinch resistor is non linear, has a low (e.g. 6-Volts) breakdown and a large variation (about 10:1), but you can cram 100k Ohm of resistance into the space of a transistor.

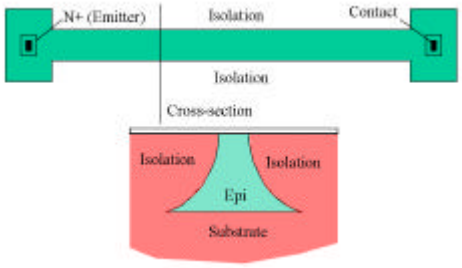


Fig. 10-25: Bulk or Epi-Pinch Resistor.

The second device is the **epi-pinch resistor**. The cross-section of a long and narrow epi region is further reduced by moving the isolation diffusions on either side very close together. Since the epi region usually is of fairly high resistivity, a substantial depletion region extends into the remaining epi region, pinching it off at an operating voltage (above the substrate potential) in excess of about 5 Volts. Thus, at any voltage higher than that, the epi-pinch resistor becomes a current source. The variation of this current is high (8:1), but you can create a small current (a few micro-amperes) in relatively little space.

Capacitors

The oxide insulating the metal interconnection from the silicon (or between metal layers) is dimensioned to give minimum stray capacitance. Even a small capacitor (say 5pF) would take up an enormous amount of space. Enormous at least in microelectronic dimensions.

Thus fabricators often provide an additional mask step to outline an area where the oxide (or nitride) is thinned down considerably, producing a higher capacitance (about 2fF/um^2 - that's femto-Farads, or 10^{-15}F/um^2). With this figure (which of course varies from process to process) a 50x50um area gives you all of 5pF, easily be the most expensive component in your chip. If you specify anything greater than 100pF, your colleagues may think you have a degree in macroeconomics.

One plate of the capacitor is always either metal or poly. For the second plate you could use a diffusion, but that creates a slight voltage dependence (there is always a depletion layer in silicon which widens as the voltage increases, adding to the distance between the plates). Poly or metal for the second plate are better choices.

The oxide underneath an MOS gate is already thinned down to achieve a reasonable transconductance, so it too has a higher capacitance

per unit area than the ordinary (field) oxide. But be careful here. At zero (DC) voltage there is no channel (source and drain form the lower plate, the gate the upper one), so the only capacitance is the one from the gate to the overlapping parts of drain and source. When the voltage exceeds the threshold, the channel comes into existence and the capacitance increases markedly. Figure 1-26 shown here depicts the behavior of a large (10x20um) 3V n-channel device.

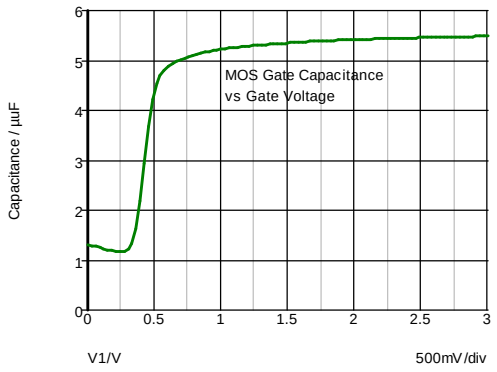


Fig: 1-26: The gate capacitance of an MOS transistor is greatly dependent on voltage.

There is also junction capacitance, which you should not dismiss lightly. The capacitance of a collector-base junction per unit area competes quite well with that of an oxide capacitor, but is voltage dependent (though not as much as the MOS gate capacitance) and the stray capacitance for one plate (collector to substrate) is higher. An even higher capacitance per unit area is offered by the base-emitter junction, though its breakdown is limited (about 6 Volts). The advantage of the junction capacitor is the elimination of the additional mask step.

Other Processes

What we have considered so far are two simple, basic processes, requiring as few as 8 masks. There are many variations, all based on these two:

- "Mixed Mode" CMOS, with devices for (somewhat) higher operating voltages and additional poly (and metal) layers;
- BICMOS processes which add full-fledged bipolar transistors to CMOS;
- Bipolar processes with vertical (high-speed) PNP transistors;
- CMOS processes with some high-voltage devices (500V).

All of these variations have one factor in common: they increase the number of masks (and processing steps) required and are thus more expensive. However, they tend to make the design of high-performance analog circuits easier, especially when both CMOS and bipolar transistors are available.

CMOS vs. Bipolar

The debate as to which is better for analog design is as old as the devices themselves. Let's examine some of the main points:

- The bipolar transistor requires an input (base) current, the CMOS devices does not. This is strictly true only at DC; at higher frequencies there is the input capacitance, which does result in a current. Also, some analog designs (see chapter 8) manage to bring this current down to a very low level.
- Bipolar transistors have lower offset voltages. Generally true, but offset voltage depends on size. Make a CMOS transistor larger than a bipolar one (or use trimming) to achieve equally low offset voltage.
- Bipolar transistors have lower noise. Again generally true, especially at low frequency (1/f noise, see chapter 6). One exception: the auto-zero (or chopper stabilized input, see chapter 8).
- CMOS devices have smaller dimensions. Generally not true. To get the required performance in an analog design (matching, gain, low noise), CMOS transistors need to be much larger than the minimum dimensions of the process would allow. At reasonably high supply voltages (3 Volts and above) CMOS and bipolar devices end up about equal in size.
- Bipolar transistors are better for low-voltage design. True. Transconductance in a CMOS device increases as the square of the gate voltage above the threshold. If the gate voltage can only go, say, 0.5 Volts above the threshold, it takes a painfully large gate-width to get a substantial drain current. In the bipolar transistor a ten-fold increase in collector current is obtained with only a 60mV (at room temperature) increase in base voltage. It is ironic that CMOS is marching toward lower and lower voltages, where it is at a serious disadvantage.

2 Simulation

In 1972 the Electrical Engineering Department of the University of California at Berkeley released the first version of SPICE (Simulation Program with Integrated Circuit Emphasis). Donald Pederson, the head of the department decided to do this free of charge; after many additions, revisions and improvements (done by a "cast of thousands" of graduate students) it is still free today.

The Berkeley SPICE program (originally written in Fortran) has been modified and sold by dozens of companies under various names. Some of the modifications were useful (such as the adaptation to PC use), many others merely served to make these programs incompatible with each other.

So, be aware that there are differences in capabilities and notation between Spice programs. Also, it is no longer true that such analysis programs running on more expensive workstations under Unix are better or faster; some PC programs (notably Simetrix) have outdistanced their Unix cousins in both speed and added features.

Simulation for analog ICs differs greatly from any kind of digital simulation. The most important factor in the latter is speed. This has led to an ever finer representation of internal capacitances and other stray-effects in the models used. In an analog IC, speed is just one of many requirements. We rely heavily on matching, and need to know the effect of the variations of many parameters in an almost unlimited number of combinations with great certainty. Each device also needs to be represented accurately over the entire operating range, not just in two states. The models, therefore, become the most important factor.

Unfortunately, the quality of models for analog or mixed-mode ICs varies greatly. Some - few - are very accurate and from simulations alone you can tell with great certainty how well your design will work in silicon, down to the exact distribution of each circuit parameter in production. But most models issued by foundries are not in this category, lacking information crucial to analog design.

In the second half of this chapter there is a fairly detailed discussion of device models for Spice. This is a somewhat tedious task, but necessary

to judge the quality of the models available to you. Read this part lightly and then use it for later reference.

What Can You Simulate?

A good analog simulator can tell you all you need to know about a design. But be aware that simulators which fall into the average category - including the most popular one - lack several of the most desirable features.

In Spice there are three basic simulations:

DC Analysis

Let's use a simple example, a buffer, a very simple circuit with a voltage gain of one. Two NPN transistors (Q1, Q2) form a differential stage, Q3 is a current mirror and Q4 an emitter follower (more about this in the next two chapters). For the current mirror we use a lateral PNP transistor with a split collector (see chapter 1).

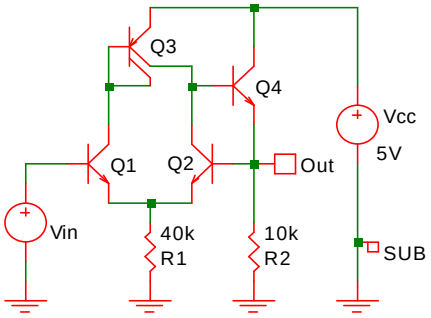


Fig. 2-1: A simple example for simulation, a bipolar buffer.

In the first DC analysis we continuously change (sweep) Vin from zero to 5 Volts and observe the output. The simulator tells us that the output follows the input, but only above about 0.6 Volts and below 4.1 Volts (the common-mode range).

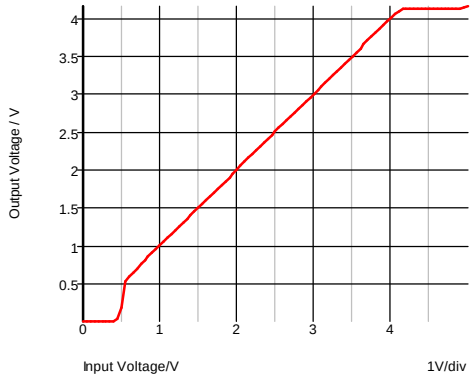


Fig. 2-2: A DC analysis, showing the common-mode range.

You could enhance this analysis by repeating it at various temperatures, i.e. by automatically "stepping" the temperature, either at regular intervals or at three or four points. While you do all this you can measure the input current (either at the base of Q1 or at either terminal of Vin), the current

consumption (at one of the terminals of Vcc), the substrate current (out of the symbol SUB) and even the power dissipation of the entire circuit or any component.

Place a current source from Out to ground and you can determine how well the circuit handles a load, i.e. determine the output impedance.

There are two sub-categories in a DC Analysis. The **Transfer Function** gives you the relationship between two nodes (not used very often) and the **Sensitivity Analysis** tells you which parameters (including transistor parameters) are most responsible for a change in a particular voltage or current at any node.

AC Analysis

The one thing you never want to forget about a Spice AC analysis is this: The signal is treated as if it were insignificantly small. You may specify a 1-Volt input signal (most people do, it represents zero dB and is thus very convenient), but the analysis program will process it without disturbing any of the bias levels. If you have a high gain, say 60dB, the output plot will show a voltage of 1000 Volts without even blushing. What it is intended to show you is the gain relative to the input; the actual values taken out of context are often absurd.

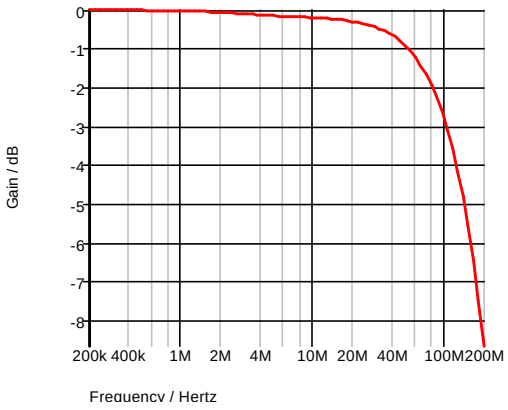


Fig. 2-3: AC analysis: gain (in dB) vs. frequency.

Our plot shows the output response of our buffer in the most simple AC analysis: a 1-Volt ac signal at the input on top of a DC Voltage of 2V, with the frequency swept from 200kHz to 200MHz. The output is in dB relative to the input, so 0dB is a "gain" of 1, -3dB is a "gain" of 0.708 (or a loss of 29.2%).

We could also move the AC source into the Vcc supply (make sure there is only one AC source per circuit). If we do this we can measure how much of a supply's ripple gets into the output, i.e. power supply rejection.

With equal ease you can measure the AC response of an output current relative to an input current. But when it comes to the relationship between a voltage and a current a measurement in dB makes little sense.

Spice also lets you measure the **phase** of any voltage or current (relative to the phase of an input signal). This is of particular interest in circuits which employ feedback, but more of this in chapters 6 and 8.

Remember though, this is a small-signal analysis, done at one particular bias point. The AC response (particularly the phase) may be different as a real-life signal moves operating voltages and currents.

An adjunct to AC Analysis is the Noise Analysis. Here the AC source is turned off and the combined effect of all noise sources inside the circuit (resistors, currents) at the output is displayed. The measure is nanovolts (or microvolts) per root-Hertz. Despite the awkward name, it is in fact an elegant measure. To get the actual noise (usually in uVrms) you simply multiply the value taken from the curve by the square-root of the frequency interval of interest. For example, between 100Hz and 1kHz we read an average of about 12nV/rtHz. Multiply this by the square root of 900 and you get 360nVrms of noise, *if* you look at it with filter which cuts out everything below 100Hz and above 1kHz. Similarly, in the flat (white noise) region between 10kHz and 1MHz we would measure about 24uVrms; even though the curve has a lower value, the total noise is much larger because of the wider frequency range.

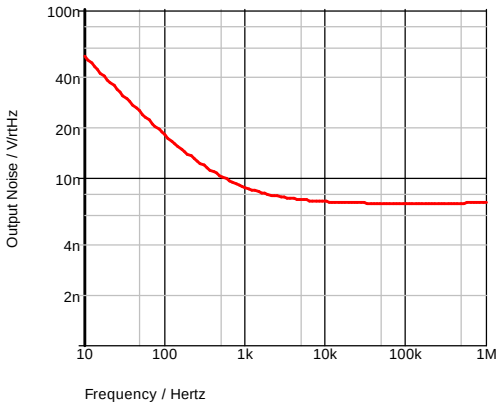


Fig. 2-4: Output noise vs. frequency.

Transient Analysis

Here we convert Vin to a pulse source (instead of DC or AC) and look at the output not over a voltage or frequency range, but time. You may have to make a few trial runs to get the appropriate pulse-width and total analysis time. At first the program will choose its own time steps, shortening the intervals when a lot of changes happen and lengthening them when no changes are taking place. But, if you are not satisfied with the resolution, you can dictate what maximum (or minimum) time-step it is allowed to take.

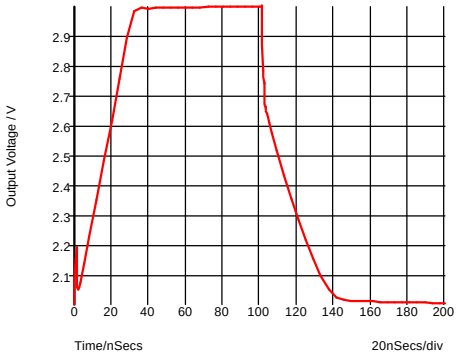


Fig. 2-5: Transient analysis 1: a 1-Volt pulse at the input.

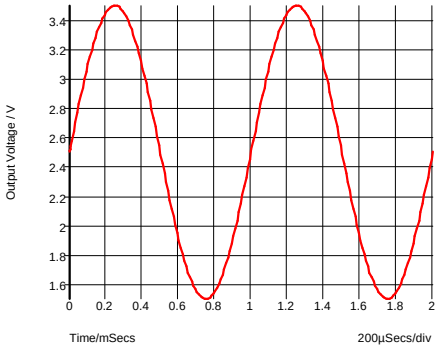


Fig. 2-6: Transient analysis 2: a 1Vpp sine-wave at the input.

Change the input to a sine-wave and you will learn a great deal more about the circuit. It will be immediately obvious if the circuit can reproduce the waveform without clipping it at either the high or low excursions. But that is only a rough impression of fidelity. What you need to know is the amount of **distortion** in the waveform.

In some programs you simply display the sine-wave, click on distortion and get the result. But if you want to have the entire information, nothing beats a **Fourier Analysis**.

The Fast Fourier Transform (FFT) is a routine which extracts the frequency components from a waveform. It is rather tricky to use and sometimes produces errors. Shown here is the result of a continuous Fourier analysis (Simetrix), which is both more detailed and more reliable.

What we see in the graph is amplitude versus frequency. At 1kHz there is the fundamental frequency with an amplitude of nearly (but not quite) 1 Volt. At 2kHz is the second harmonic with an amplitude of 100uV, or 0.01%. The third harmonic measures about 12uV, or 0.0012%.

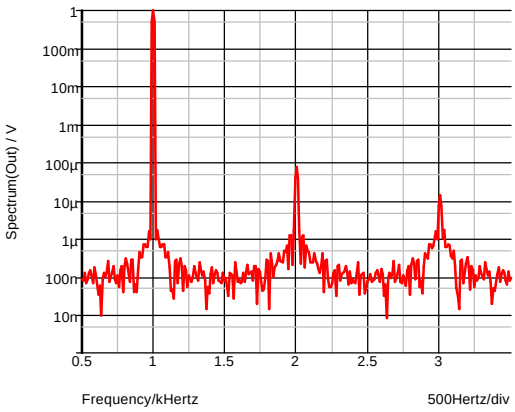


Fig. 2-7: Fourier analysis.

To get this kind of resolution you need to run the sine-wave for many cycles (at least 1000) with small enough time steps.

Noise analysis in the small-signal (AC) mode has strict limitations. It presumes that the operating voltages and currents are steady. This is fine for a circuit which is perfectly linear, but it falls down if a design is non-linear, either by design or by mishap. Take, for example, the case of a mixer (or modulator). A signal of a particular frequency enters a deliberately non-linear block, such as a diode mixer or the phase-detector of a phase-locked loop. The non-linearity creates other frequencies (usually much lower ones, such as frequency differences), one of which we use and amplify. An AC noise analysis is useless here, because it cannot follow what happens to the noise as it is transformed by the mixer.

What we need in such a case is a transient analysis program which pays attention to noise sources. Few have this capability; a notable exception (again) is Simetrix.

The Big Question of Variations

As pointed out in chapter 1, device parameters in an IC vary from run to run and from wafer to wafer. The devices are made at temperatures at which the material is no longer a semi-conductor. You have to wait for the wafer to cool down to measure the parameters of a diffusion.

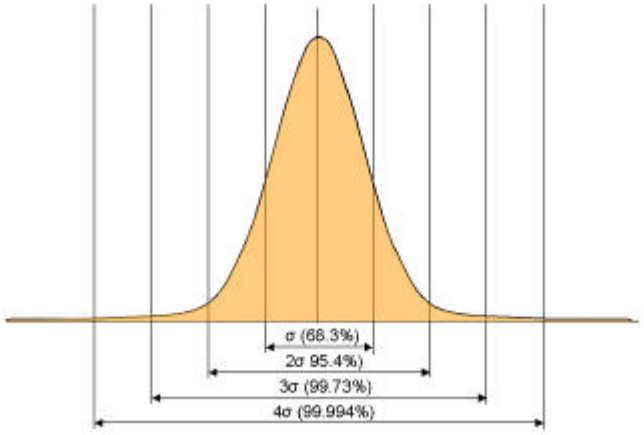


Fig. 2-8: The normal or Gaussian distribution.

Most parameters follow a "normal" (Gaussian) distribution. There is a mean value, at which the number of occurrences is maximum. A deviation of ±s (sigma) from this point contains 68.3% of all measured

values. If you allow the deviation to be three times as large (±3s) you enclose 99.73% of all measurements.

This sounds like you are discarding only 0.27% of all values, but the figure is deceiving. So far we have considered one parameter only, but there are many in an integrated circuit. Suppose your design is influenced by 50 of them. The total parameter "yield" then is $.9973^{50}$, or 87.4%. In other words, you would discard 12.6% of all chips on a wafer.

Unless you simply don't care about cost, you need to design an analog IC so no chips are lost because of parameter variations, i.e. the design can withstand a variation of each and every device parameter to *at least* 3σ; 4σ would be better.

But how do you find out how much parameter variation your design can take? The answer is **Monte Carlo analysis**, and only Monte Carlo analysis.

There is in use what is called a "four-corner analysis". Device parameters are bundled together in four groups, representing extremes, or worst cases. The plain fact is this: it doesn't work for analog circuits. The four-corner models are just barely able to predict the fastest or slowest speed of digital ICs, but the grouping simply doesn't apply to analog ones. In fact, no grouping is possible; a parameter's influence differs from design to design. Analog designers who are satisfied with a four-corner analysis simply fool themselves into believing that they have a handle on variations, when in truth the result is quite meaningless.

A true Monte Carlo analysis varies the device parameters in a random fashion, so that *every combination of variations* is covered. This is also what you get in production.

You don't need to vary every parameter of a device, only the major ones. For example, varying IS, BF and the capacitances in a bipolar transistor model is sufficient; the same is true for the threshold voltage, the transconductance and the capacitances in MOS transistors. If matching is expected, there must be two additional entries, one for the absolute variation, and one for the variation between devices on the same chip. These "tolerances" are either inserted into the model file directly or contained in a separate file, depending on the analysis program used. The Monte Carlo program then simply runs the chosen analysis repeatedly, each time with a different set of variations, randomly chosen. Our example shows the variation over temperature for a bandgap reference (untrimmed).

How many runs need to be specified for a Monte Carlo analysis? There is an easy way to find out. Start with 20. Then increase this number until the extremes no longer change. For this analysis 50 runs were used,

which is more than needed. But with today's fast computers you can afford to go overboard: the analysis took all of 8 seconds!

This one picture gives you the variation of a reference voltage over temperature, *exactly as it will happen in production*. You notice a Gaussian distribution (more curves at the center, fewer at the extremes). Between the top and bottom curves lie 99.73% (3-sigma) of all circuits.

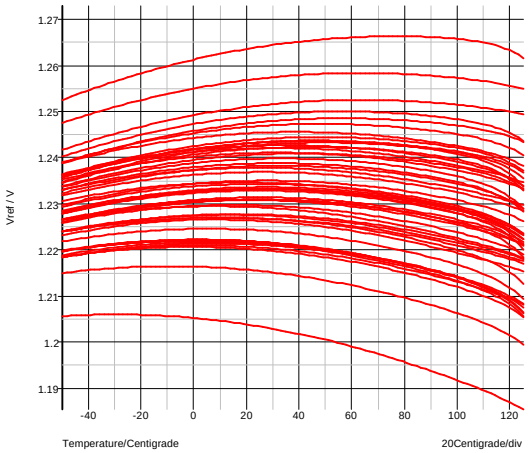


Fig. 2-9: The result of a Monte Carlo analysis. Each curve represents the behavior of one circuit (among 50) in production.

Without the Monte Carlo analysis we would not know how much variation to expect until *several* wafers from *several* different lots have been tested (a single prototype wafer cannot tell you what the variations in production will be).

Models

The Diode Model

As we have seen in chapter 1, there isn't any one junction in an IC which can be used directly as a diode; a "diode-connected" transistor does this job with greater accuracy and far fewer side-effects.

However, a bipolar transistor consists of junctions, at least two of them. Thus a model for a junction diode is a fundamental element in models, even in CMOS. In the model file (which is always in ASCII) you might see the following:

.MODEL Diode1 D IS=1E-17 RS=20 CJO=0.85E-12

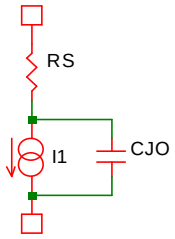


Fig. 2-10: a simple diode equivalent circuit.

A model statement always starts with a dot. "Diode1" is the name of the device (which can be anything) and

D says the device is a diode. The remaining entries are in Amperes, Ohms and Farads.

This is about as simple a diode model as you can possibly make it; just three parameters are specified. IS, together with the series resistance RS, determine the DC characteristics, CJO the junction capacitance. Let's first look at the DC behavior.

As we have seen in chapter 1, the current/voltage relationship of an ideal junction is given by:

$$I1 = Is\, e^{\frac{Vd \cdot q}{k \cdot T}}$$

where Is = the diffusion current
Vd = the voltage across the current source (i.e. not including RS)
q = the electron charge
k = the Boltzmann constant
T = the temperature in Kelvin

In Spice this diode equation is greatly expanded:

$$I1 = IS\left(e^{\frac{Vd \cdot q}{N \cdot k \cdot T}}\right) \cdot e^{(\Delta T - 1) \cdot \frac{EG \cdot q}{N \cdot k \cdot T}} \cdot \Delta T \frac{XTI}{N} + ISR\left(e^{\frac{Vd \cdot q}{NR \cdot k \cdot T}}\right) \cdot \left(\left(1 - \frac{Vd}{VJ}\right)^2 + 0.005^{\frac{M}{2}}\right)$$

At first this equation might appear utterly complicated, but it isn't. If you look at the first portion (to the left of the + sign), you see three multiplied constants:

N = the Forward Emission Coefficient (1)
EG = the energy gap, which depends on the material (1.11 for silicon); you set this to 0.69 for a Schottky diode, 0.67 for germanium and 1.43 for Gallium-arsenide.
XTI = the temperature coefficient of IS (3).
ΔT = the quotient of operating (i.e. junction) temperature to room temperature (usually 300K or about 27^{o}C).

With these three constant you can shape the basic exponential curve to what is actually measured. If you don't list them in the model statement, they assume the values shown in parentheses.

The portion of the equation to the right of the + sign adds a leakage current, i.e. a small current in excess of the (reverse) current predicted by the ideal diode equation. You can modify the shape of its curve with the constants NR, M and VJ.

As shown, the model makes the breakdown voltage of the diode infinite. You can limit this with the parameter BV and three companions:

TBV1 (its first-order temperature coefficient)

TBV2 (second-order temperature coefficient)
IBV (the current at which breakdown is specified)

There is also a parameter, IKF, which splits the DC curve into two regions. In ICs this is very rarely used.

The series resistance, RS, which also influences the DC behavior, has first and second-order temperature coefficients, TRS1 and TRS2.

The junction capacitance shown in the model, CJ0 (or CJO) is measured at zero voltage. Since its value at different voltages (forward or reverse) depends on the grading of the junction (abrupt, diffused, implanted etc.), it too is modified by three constants, VJ (1), M (0.5) and FC (0.5). If you don't list them in the model statement, the default values in parentheses will be used.

For a voltage across the diode (not including RS) equal to or less than the product of FC and VJ, the formula is:

$$C = CJO \cdot \left(1 - \frac{Vd}{VJ}\right)^{-M}$$

If the voltage across the diode is greater then FCxVJ:

$$C = CJO \cdot (1 - FC)^{-(1+M)} \cdot \left(1 - FC \cdot (1 + M) + M \cdot \frac{Vd}{VJ}\right)$$

There are two noise sources in a diode: the resistor RS and the current I1. Without any additional parameters these are treated as white noise sources, i.e. the noise is the same at any frequency. (For a more detailed look at noise see chapter 6). Since there is also flicker noise (which increases at low frequency), two constants, KF and AF are used and the following expression is added to the current source noise:

$$\frac{KF \cdot I1^{AF}}{f}$$

The Bipolar Transistor Model

42 parameters are used to represent a bipolar transistor in Spice. While the number may look a bit daunting, it is actually quite straight-forward once you are familiar with the Spice diode, and they are placed into five groups:

Base-Emitter Diode. Here we have the plain diode Spice parameters as discussed above, but re-named; the abbreviations in parentheses refer to the ordinary diode parameters: IS (IS), NF (N), ISE (ISR), NE (NR), RE (RS), EG, XTI, CJE (CJO), VJE (VJ), MJE (M), FC. The series resistance of the diode is divided into two parts: RE (at the emitter end, with the emitter current flowing through it) and RB (at the base-contact end). The latter

starts at RB at low current and drops gradually to the value RBM at a current specified by IRB; this reflects the use of the entire emitter at low current and only the emitter edge facing the base contact at high current, as discussed in chapter 1.

Current Gain: The main parameter here is BF (the forward beta, or hFE) and its temperature coefficient XTB. Without any additional parameter the current gain would be the same at any collector voltage or current. The Early effect is represented by VAF (the Early voltage, see chapter 1). The drop-off at high current is produced by IKF (the current at which hFE starts to drop) and NK (the steepness of the drop). ISE and NE of the base-emitter diode are responsible for the drop in hFE at the low-current end; simply shunting a small amount of base current to the emitter.

Reverse Current Gain: You may be convinced that you will never operate a transistor with the collector and emitter interchanged, but just in case provided the parameters BR, NR, VAR, IKR and TR .

Base-Collector Diode: Here again we have the basic diode Spice parameters, again renamed: ISC (IS), NC (N), RC (RS), XTF (XTI), CJC (CJO), VJC (VJ), MJC (M) and TF (TT). The last one is the transit time (now through the base to the collector) which accounts for any delay which cannot represented by capacitance alone; it is embellished by ITF (which makes TF dependent on current), VTF (showing dependence of TF on base-collector voltage) and PTF (an excess phase at a frequency $1/(TF \times 2\pi)$).

Noise: As in a simple diode, additional low-frequency (flicker) noise is represented by the parameters KF and AF, but here they work on the collector current.

The Spice model for an integrated bipolar transistor can have either three or four terminals. The fourth terminal (of an NPN transistor) is the substrate and between it and the collector there is a diode, represented by the five parameters ISS, NS, CJS, VJS and MJS. This is a major flaw in Spice, for a mere diode here is inadequate. When the transistor saturates, a substantial portion of the total current flows to the substrate, which this model simply ignores.

Fortunately Spice also contains a solution to this problem. To represent an NPN transistor correctly, you need to add a second transistor; Spice lets you combine the two (or any number of devices) in a **subcircuit**.

When N1 saturates (i.e. the collector drops below the base potential), the PNP transistor P1 becomes active and draws base current to the substrate. This is what happens in real life: there is a stray PNP transistor, formed by the base (P), the collector (i.e. the epitaxial layer, N) and the substrate (P).

Rather than put the stray capacitance (and leakage current) between collector and substrate into the PNP transistor (which is somewhat cumbersome) a separate diode DCS is inserted. DZ, a Zener diode, corrects another flaw of Spice: there are no breakdown voltages in the bipolar transistor model; for the base-emitter diode we need this effect, it is sometimes used as a Zener diode. If you also want to have a collector-emitter breakdown, place an additional Zener diode between collector (cathode) and base.

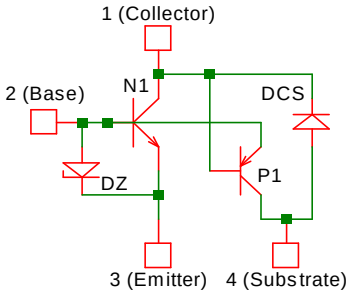


Fig. 2-11: Equivalent circuit for an integrated NPN transistor.

The model for this subcircuit looks as follows:

```
.SUBCKT NPN1 1 2 3 4
Q1 1 2 3 N1
Q2 4 1 2 P1
D1 2 3 DZ
D2 4 1 DCS
.ENDS
```

The first line, after the .SUBCKT (all models start with a dot) lists the name of the subcircuit and the order of connections (which will be followed in the netlist). The next four lines list the device types, the connections and the name of the model. The last line signified the end of the subcircuit listing.

Spice lets you define **global nodes**. This is especially convenient for the substrate and avoids cluttering up the schematic with unnecessary lines. This feature is used throughout this book for bipolar devices. However, you will have to remember to place the contact to the substrate (SUB) at the appropriate point (almost always the most negative supply).

Spice now needs a model for each of the devices used. For example (for a 20-Volt process):

```
.MODEL N1 NPN IS=3.8E-16 BF=220 BR=0.7
+ ISE=1.8E-16 IKF=2.5E-2 NK=0.75 IKR=3E-2 NE=1.4 VAF=60
+ VAR=7 RC=63.4 RB=300 RE=19.7 XTB=1.17 XTI=5.4
+ TF=1.5E-10 TR=6E-9 XTF=0.3 VTF=6 ITF=5E-5 CJE=0.21E-12
```



```
+ MJE=0.33 VJE=0.7 ISC=5E-12 KF=2E-13 AF=1.4
.MODEL P1 PNP IS=1E-15 BF=100 CJE=0.175E-12 XTI=5.4
+ MJE=0.38 VJE=0.6
.MODEL DZ D IS=1E-18 RS=250 BV=5.9 IBV=10UA
+ TBV1=1.8E-4
.MODEL DCS D IS=1E-17 RS=10 ISR=5E-12 CJO=0.85E-12
+ M=0.42 VJ=0.6
```

Note that the model for DZ has no capacitance; this is already present in the base-emitter diode of N1.

The Model for the Lateral PNP Transistor

For a lateral PNP transistor the Spice bipolar transistor model alone is woefully inadequate. This type of transistor not only produces a substrate current when it saturates but also in its normal operation; neither of these is present in the Spice model.

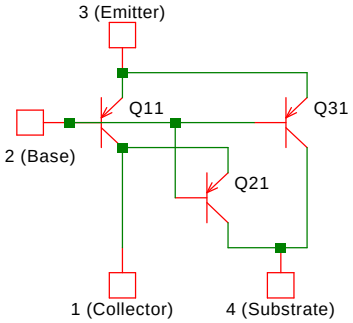


Fig. 2-12: Equivalent circuit for a lateral PNP transistor.

To correct this flaw, we need to use a subcircuit again, only this time two additional transistors are required, one to cause the substrate current at saturation (Q21) and one at normal operation (Q31); the parameters of the latter (particularly IS and BF) are chosen so that the substrate current is smaller than that of Q11 (generally about 20%).

The model for this subcircuit looks like this:

```
.SUBCKT PNP1 1 2 3 4
QP11 1 2 3 QP1
QP21 4 2 1 QP2
QP31 4 2 3 QP3
.ENDS
```

And the models, again for an arbitrary example of a 20-Volt process:

```
.MODEL QP1 PNP IS=1E-16 BF=89 VAF=35
+ IKF=1.2E-4 NK=0.58 ISE=3.4E-15 NE=1.6 BR=5
+ RE=100 RC=800 KF=1E-12 AF=1.2 XTI=5 ISC=1E-12
+ CJE=0.033E-12 MJE=0.31 VJE=0.75 CJC=0.175E-12
+ MJC=0.38 VJC=0.6 TF=5E-8 TR=5E-8
+ XTF=.35 ITF=1.1E-4 VTF=4 XTB=2.3E-1
.MODEL QP2 PNP IS=5E-15 BF=150 RE=100 TF=5E-8 XTI=5
```



```
.MODEL QP3 PNP IS=1E-18 BF=25 CJC=0.85E-12
+ MJC=0.42 VJC=0.6 XTI=5 RE=100
```

MOS Transistor Models

Once upon a time there was a company which brought out its own variation of the Berkeley Spice program: HSPICE. It specialized in making refined models for MOS transistors, many of them. The models were called levels and many companies bought their own levels, like boxes at the opera. AMD had three of them, Siemens acquired two. Motorola, National Semiconductor, Sharp, Cypress, Siliconix and a few others only got one. By 1995 there were 39 such levels and us poor ordinary folks couldn't access most of them: they could only be used by the company who had sponsored them.

At last good old Berkeley came to the rescue. A team of researchers developed the **BSIM** model (Berkeley Short-channel IGFET Model). The team stayed with it, through BSIM1, BSIM2, BSIM3 and even BSIM4. These models divided the MOS transistors in ever finer structures, tracking the trend toward geometries far below 1um. As of this writing BSIM3.3 is the dominant model in the industry, leaving the many HSPICE levels in the dust.

Naturally HSPICE took the BSIM models and made its own version, adding more levels.

The increasing BSIM refinements have its toll: the number of parameters has become very large, so large that it takes an entire book to explain them. For digital ICs, which require utmost speed, this simply has to be accepted. For analog designs, which invariably use larger dimensions to obtain adequate performance (especially for matching), it is a burden only grudgingly tolerated. MOS model-making has become an art dominated by the digital realm, of limited use to the analog designer.

In a modern BSIM model you are confronted by a mass of data which almost always is presented in an arbitrary way, lacking an organization which would make it more understandable. To help in a minor way, they are grouped here; the bold-faced parameters are absolute values; all others are modifiers. Parameters in square brackets are temperature coefficients.

Threshold Voltage: **VTHO**, K1, [KT1, KT1L], K2, [KT2], K3, K3B, DVT0, DVT0W, DVT1, DVT1W, DVT2, DVT2W, VBM, VOFF, KETA, PSCBE1, PSCBE2.

Mobility: **UO**, UA, [UA1], UB, [UB1], UC, [UC1].

Saturation: **VSAT**, [AT], A0, AGS, A1, A2, B0, B1, DELTA, EM, PCLM, PDIBLC1, PDIBLC2, PDIBLCB, DROUT, PVAG, AGS, ALPHA0, BETA0.

Sub-Threshold: ETA0, ETAB, NFACTOR, DSUB.

Geometry: **W0**, DWB, DWG, LL, LLN, LW, LWL, LWN, WL, WLN, WW, WWL, WWN.

Capacitances: **CGS0**, **CGD0**, **CGB0**, **CJ**, MJ, MJSW, PBSW, **CJSW**, MJSW, CJSWG, MJSWG, PBSWG, PB, **CGSL**, **CGDL**, CKAPPA, **CF**, CLC, CLE, DLC, DWC, ELM, **CDSC**, **CDSB**, **CDSCD**, **CIT**,

Resistances: **RSH**, **RDSW**, [PRT], PRWB, PRWG, WR, LINT, WINT

Process Parameters: **TOX, XJ, XT, NCH (PCH), NGATE, NLX, NSUB, GAMMA1, GAMMA2**, **JS**, [XTI], NJ, **JSSW**.

Noise: AF, KF, EF, EM, NOIA, NOIB, NOIC

BSIM models also allow "binning": several models are written for different geometries of the same device, and then selected to fit into a range of gate width and length with the parameters LMIN, LMAX, WMIN and WMAX. While this is not really necessary for the parameters listed above (some foundries, notably AMS, manage to create equally accurate model without binning), the Monte Carlo variations should be tied to channel width and length (i.e. area). Note that the multiplier M is used for transistors with a channel width beyond WMAX.

To get into more detail on the many parameters, you will need to consult the original Berkeley documentation (see references). Be forewarned: this is a lengthy document.

Resistor Models

A Spice resistor model has no stray capacitance, nor does it recognize any possible effect by surrounding layers. There are some cases where such a simple model is inadequate. For example, the frequency (and phase) response of large-value resistors (50kΩ and more) can be significant enough to bring about oscillation in a feedback path. Also, an error is introduced in a divider, if the resistors are diffused and placed in the same pocket (or "tub"); each resistor is at a different DC potential and their voltage dependence will result in slightly different values. This error becomes large with ion-implanted resistors.

Some simulation programs have the capability to extract stray capacitances from the layout, but few pay heed to voltage dependence. If you want the complete behavior *before* the layout is done, here is a model:

```
.SUBCKT RCV 1 2
R1 1 4 RB {m/3}
R2 4 5 RB {m/3}
R3 5 6 RB {m/3}
V1 6 2 0
B1 6 1 I=I(V1)*(0.0033*((V(3)-(V(1)+V(2))/2))^0.6)
D1 1 3 DRSUB {m/2}
D2 4 3 DRSUB {m}
D3 5 3 DRSUB {m}
D4 6 3 DRSUB {m/2}
.ENDS
.MODEL DRSUB D IS=1E-16 RS=50
+ CJO=2.7E-14 M=0.38 VJ=0.6
```

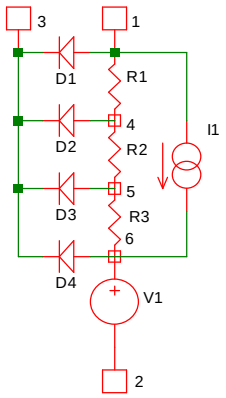


Fig. 2-13: Equivalent circuit for a 3-segment integrated resistor.

This is again a subcircuit. The resistor is divided into three equal sections and the stray capacitance is represented by four diodes to the surrounding n-type material (assuming that the resistor is p-type).

To model the voltage dependence, the current is measured in the dummy voltage source V1 (with zero voltage) and from it a current I1 is created and subtracted from the total current through the resistor. The value of this current is:

$$I1 = I(V1)\cdot\left(0.0033\cdot\left(V(3)-\frac{V(1)+V(2)}{2}\right)\right)^{0.6}$$

where 0.0033 and 0.6 determine the amount and shape of voltage dependence. Note that the bias voltage is applied from terminal 3 to the

mid-point of the resistor. B (in Simetrix) is an arbitrary function, serving as a current-controlled current source.

Contrary to common belief, a three section lumped model is remarkably accurate. Compare the frequency response of such a model with one that has 160 sections.

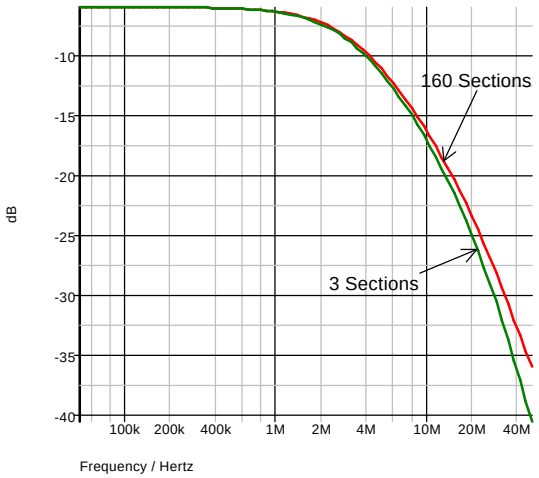


Fig. 2-14: Comparison of lumped resistor models.

Models for Capacitors

There are only two cases where a simple capacitor model (i.e. an ideal capacitor) is inadequate:

1. There is a requirement for unusual precision. If one plate of an oxide capacitor is a diffused layer (or a poly layer with a high sheet resistance) the capacitance will decrease slightly as the potential across the plates is increased. A competent model will reflect this non-linearity.

2. The capacitor is used at the high-frequency end. Here it is not of great importance for the model to show the non-linearity, but to reflect any series resistance and stray capacitances from both the lower and the upper plate to neighboring regions.

Pads and Pins

If you are working at high frequencies - say above 50MHz - you need to consider the properties of the pads, the ESD protection devices, the bonding wires and the package pins. A pad has a capacitance to the underlying layer (usually ground); with an ESD protection device this can easily amount to more than 1pF. The bonding wire has an inductance; it may be small (perhaps 7nH, but this depends on the length of the wire), but it begins to play a role above about 100MHz. Then there is the package pin capacitance, which is not to ground but between pins (about 1pF, but greatly dependent of the package).

Just How Accurate is a Model?

The quality of device models from wafer-fabs varies greatly. A few are outstanding, unerringly accurate and complete. Others are so bad that they will almost guarantee major flaws in your design. The majority of them are incomplete for analog design.

It pays to examine the models before starting to simulate. If the NPN transistor model is not a subcircuit, use it with caution; behavior in saturation is going to be different in the real circuit. If the lateral PNP model is not a subcircuit, it doesn't make much sense to use it at all.

A set of device models is not really ready for use until it has been tested in actual circuits. Unfortunately models are commonly put together by people who are not designers (especially not analog designers), so they tend not to be verified in real-world applications.

This is especially true for bandgap references (see chapter 7), which demand uncommon accuracy from the bipolar transistor models. Even a small error in VBE (i.e. the basic diode voltage) and its temperature coefficient causes intolerable errors in the reference voltage. Here it is in fact preferable to set such parameters as IS (and its modifiers) so that it fits (several) designs existing in silicon.

You should also check the models for the presence of Monte Carlo parameters. If there aren't any, you are going to be seriously handicapped for an analog design.

3 Current Mirrors

Bob Widlar was a truly great designer of analog ICs. He was wild and totally unmanageable and had an odd sense of humor. The press loved him and he had a flair for self-promotion. He shunned computer analysis, preferring to breadboard his circuits, but time and time again he came up with nuggets of design details and products which were thought to be impossible. Burned out by the frenzy of Silicon Valley he moved to Mexico, where he died in 1991 at age 53.

One of Widlar's early contribution is the current mirror, a design detail (or design element) which you will now find in just about any analog IC.

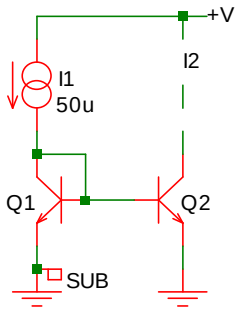


Fig. 3-1: The Widlar current mirror.

Start with the primary current, I1, which flows into the diode-connected transistor Q1. This produces a voltage drop across Q1, namely that of its base-emitter diode; this voltage drop is called a VBE.

Now connect the base and emitter of a second, identical transistor, Q2, to the same nodes as those of Q1. Since the base-emitter voltage of Q2 is the same as that of Q1, it follows that its collector current should be the same as that of Q1 and, therefore I2=I1.

Well, not so fast. There are errors, two of them. The first one concerns the base currents. I1 splits into three paths: the collector current of Q1 and the two base currents. Assuming a minimum current gain of 100, each base current amounts to 1% of the collector current, for a total of 2%. So the collector currents of Q1 and Q2 are 2% smaller than I1, worst-case.

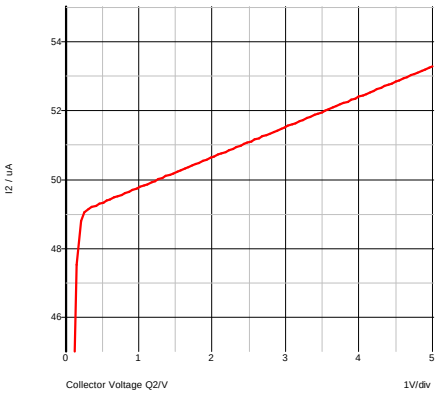


Fig. 3-2: The current of Q2 depends on its collector voltage.

The two transistors may be identical, but they are not necessarily operated identically, which is error number two. The collector voltage of Q1 is always VBE, but the collector voltage of Q2 may be anything. As

The Current Source

All current mirrors start with a current source, from which one or more currents are derived. For ICs, a current mirror is a more basic element than a current source, which is the reason they are discussed first.

However, be aware that there is a significant difference between a theoretical current source (as in a simulation) and a practical one. In a simulation a current source will do anything to keep its programmed current level, including building up thousands of volts. In an actual circuit the supply voltage limits the excursion.

Also, little distinction is usually made between a current source and a *current sink* (e.g. I1 in figure 3-3) . For convenience all of them are usually termed current sources.

we have seen in chapter 1, the gain is affected by the collector voltage (the Early effect), increasing as the collector voltage is increased. Thus I2 is not exactly steady. For this particular transistor (made in a process capable of 20 Volts) the change amounts to 8% from 0.3V (the saturation voltage of Q2) to 5 Volts. (I1 is 50uA for all examples in this chapter).

This is the most simple current mirror and, as we shall see in figures 3-7 and 3-9, we can improve its performance considerably with additional devices. There is also a lateral-PNP equivalent. Using a split collector (see figure 1-17), this current mirror needs only a single device. Each collector being smaller, the maximum current is more limited (depending on the process, about 100uA).

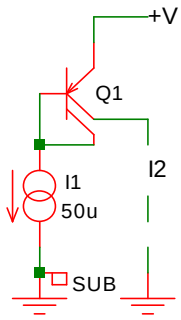


Fig. 3-3: Current mirror with lateral PNP.

The voltage dependence of a PNP current mirror is generally a bit worse than that of an NPN design (here about 12% change). The voltage of the second PNP collector can move to within about 0.3 Volts of +V. If you let it go any higher (or disconnect the collector completely), you get a substrate current about equal to I1.

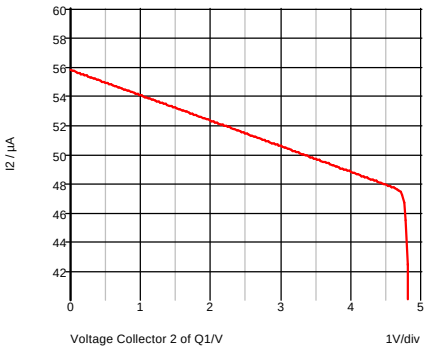


Fig. 3-4: Voltage dependence of I2.

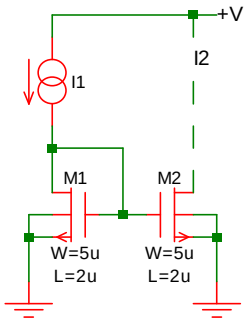


Fig. 3-5: Simple MOS current mirror.

The current mirror also works with MOS devices, but it is not quite correct to call M1 a "diode-connected transistor" (there is no "junction" diode).

The change in current is only about 1.5% from 1 to 3 Volts (0.35u process), but only because the channel lengths were made quite large. It takes at least 0.5 Volts at the drain of M2 to make the mirror work, a figure which you can improve by making the devices much wider. The mirror can be inverted by using p-channel devices.

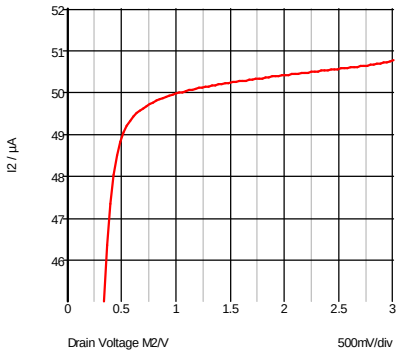


Fig. 3-6: The voltage dependence of an MOS current mirror can be made smaller by increasing channel length.

Now let's see how we can improve the performance of the basic Widlar circuit (not that we are any smarter than Widlar, but we have had a long time to work on it and have much better tools now).

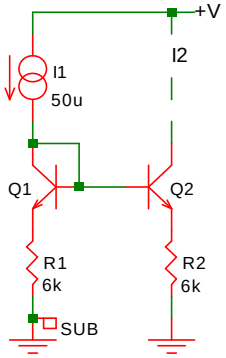


Fig. 3-7: Improved current mirror with emitter resistors.

A first step is to place resistors in the emitters (or the sources in case of MOS). With 6kOhm in the example here we drop 300mV across the resistors. If the current in Q2 wants to be higher than I1, it would also cause a higher voltage drop across R2. This latter increase forces I2 back to where it is more or less equal to I1. There is, however still the base current error, which is not improved.

There is a penalty: The voltage at the collector of Q2 cannot go any lower than the voltage drop across R2 plus the saturation voltage of Q2, about 600mV total for this case. From this point to the supply voltage (5 Volts) I2

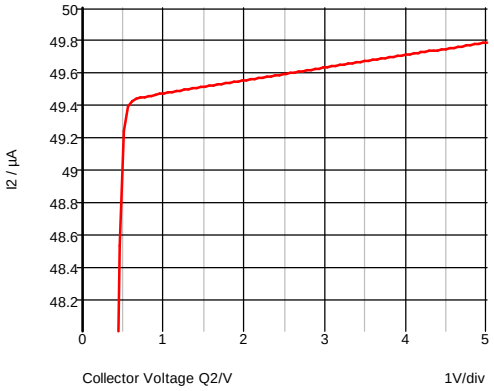


Fig. 3-8: Voltage dependence is now reduced to 0.7%.

changes only about 0.7%. A measure of quality of a current source is its output impedance, i.e. the change in voltage divided by the change in current. This has now increased from 1.1MOhm for the original current mirror to 12MOhm.

If you want to use emitter resistors in the PNP equivalent, you will need to use two separate transistors rather than a split collector.

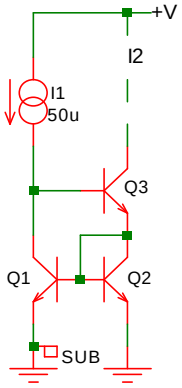


Fig. 3-9: Wilson current mirror.

An even greater improvement can be made with the addition of a transistor. This circuit, invented by George Wilson, is naturally called the **Wilson Current Mirror** (analog designers don't get Nobel prizes, they get a circuit named after them). Q3 acts as a **cascode** stage; its sole job is to shield the important matching transistors, Q1 and Q2 from any fluctuation in the output voltage. It does this job and more: by a happy coincidence the three base currents cancel and I2 is now within about 1% of I1 and changes only about 0.09% over the useful voltage range (an output impedance of 90MOhm).

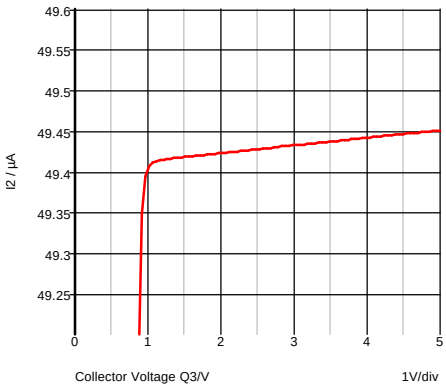


Fig. 3-10: Performance of the Wilson current mirror.

Note, however, that the useful voltage range stops at a little over 1 Volt, given by the VBE of Q2 plus the saturation voltage of Q3.

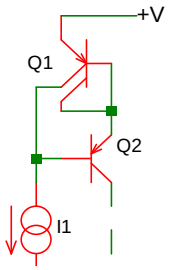


Fig. 3-11: PNP Wilson current mirror.

Naturally there is a PNP equivalent for the Wilson current mirror. We can again use a split-collector device for Q1. The output voltage can go to within about 1 Volt of +V (at room temperature). Here the improvement is not quite as good (the output current changes by about 0.5%).

There is still a systematic error in the basic Wilson current mirror: The two transistors intended to match don't have the same collector voltages; one is at VBE, the other at 2 VBE. In the relentless pursuit of perfection, given at birth to all analog designers, we shall now proceed to eliminate it. Enter a fourth transistor.

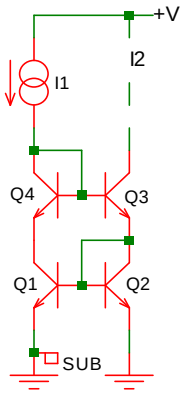


Fig. 3-12: Four-transistor mirror.

The only purpose of Q4 is to lower the collector voltage of Q1 to the same level as that of Q2. With this I2 is now within 0.6% of I1 and changes by less than 0.08% with voltage.

The single sweep in this DC analysis is, however, deceiving. The depicted curve can only be observed once in a great while, when all four devices match perfectly. Only a Monte Carlo analysis can tell you what really will happen in production. The remarkably small change with output voltage is a fact, but the output current will vary by ± 3% because of mismatch.

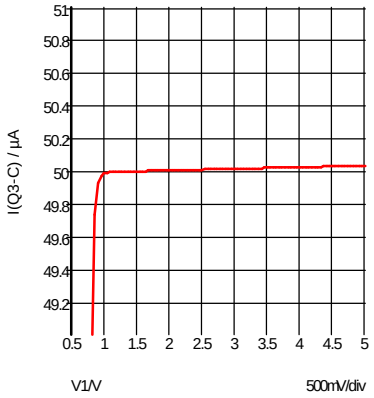


Fig. 3-13: Performance of the four-transistor current mirror.

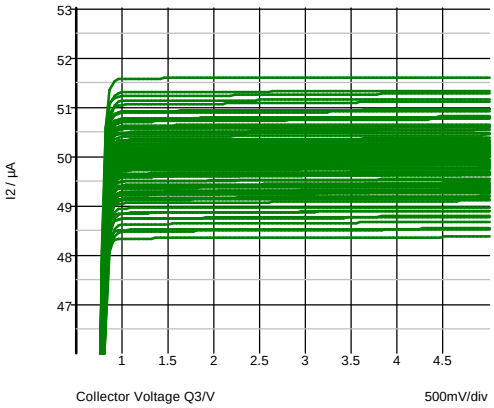


Fig. 3-14: Although the output current now changes little with voltage, there is still considerable variation due to mismatch, as a Monte Carlo analysis will show.

Current mirrors need not be restricted to 1:1 relationship between input and output current. If the critical transistor on the output side is increased in size, its collector current is increased too. In a bipolar transistor the current ratio is determined by the size of the emitter (more precisely, the active emitter length; see chapter 1) but a accurate ratio is in practice only achieved if you work with a number of identical emitters. In figure 3-15 Q1 has one emitter while Q2 has three (they can all be in the same base),

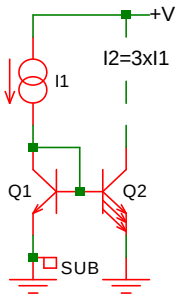


Fig. 3-15: 1:3 current ratio.

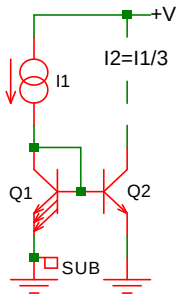


Fig. 3-16: 3:1 current ratio.

resulting in a current which is three times that of I1. In figure 3-16 Q1 has three emitters and Q1 one, which causes I2 to have one-third the value of I1. Any ratio is possible (such as 3:2 or 5:3). In a CMOS design the ratio can be

obtained simply by varying the channel width of one of the transistors, but best matching is achieved by using identical, multiple devices.

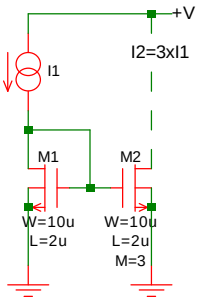


Fig. 3-17: MOS 1:3 current ratio.

This scheme can be expanded to creating multiple currents (i.e. additional transistors with their bases and emitters connected in parallel to those of Q2, but their collectors separate) in any ratio you desire. But, in bipolar circuits, there is a limit: the base current for each additional emitter is supplied by I1. Thus, with Q2 having three emitters (or two additional transistors), the systematic errors (with a minimum gain of 100) is 4%; with 9 additional emitters this increases to 10%.

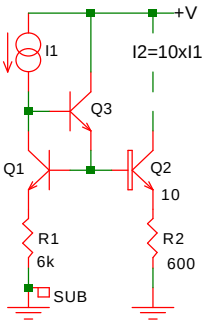


Fig. 3-18: an additional transistor supplies the base current.

There is a solution to this (have you noticed, there is always a solution, it just takes one or a few additional transistors). Here, with the help of Q3, the base current for Q1 and Q2 is supplied not from I1 but from the positive supply, thus the base current error is divided by the gain of Q3. In this way you can not only create large current ratios but also drive a substantial number of separate transistors. In Fig. 3-18 emitter resistors are used to get less of a change in I2 with a varying output voltage (0.7% from 0.7V to 5V); if you have 10 separate transistors, they all get 6kOhm in the emitter; if the current is simply multiplied by 10, R2 has one-tenth the value of R1. Remember that best matching is achieved if the resistors consist of identical sections, i.e. you create a basic 600 Ohm device and use one for Q2 and 10 in series for Q1.

If you are thinking of turning I1 on and off rapidly, be aware that this circuit is very slow to turn off; there is no discharge path for the bases of Q1 and Q2. A resistor (or another current sink) from these bases to ground helps to speed up the turn-off time.

The base current problem does not exist with CMOS devices, they require no input current. Here you are free to add as many dependent current sinks as you desire - if the change in current with output voltage doesn't bother you. If this voltage dependence is too large, you have the choice of increasing the gate lengths, adding

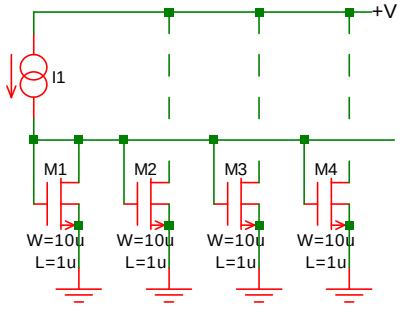


Fig. 3-19: Multiple current mirrors in MOS.

resistors in all sources or - you guessed it - add a few devices.

To reduce the influence of the output voltage we could use the Wilson current mirror, as discussed above. But MOS devices cannot take advantage of one of its features, the cancellation of base currents. For this reason the Wilson current mirror is not the best choice for MOS, a simple cascode stage has slightly better performance and can be made to have a wider range in the output voltage.

Here M3 and M5 simply shield M2 and M4 from changes in the output voltage. Their gates are held at a voltage slightly higher than the threshold voltage of M1 by causing a voltage drop in R1. In our case here this bias voltage is 500mV, which results in quite a remarkable

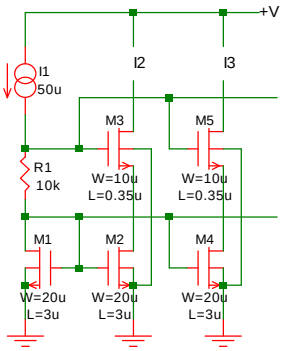


Fig. 3-20: Current mirror with cascode transistors.

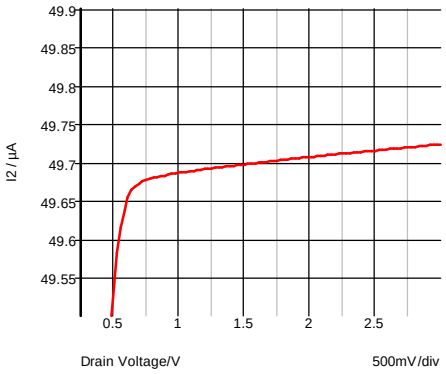


Fig. 3-21: Performance of cascode MOS current mirror.

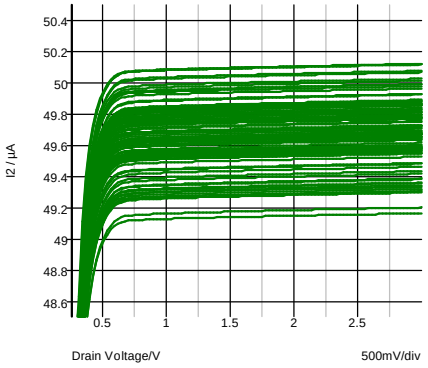


Fig. 3-22: Figure 9-21 repeated with Monte Carlo variations.

performance, but requires at least 0.7 Volts at the output. Lowering the voltage drop across R1 lowers this minimum output voltage, but increases the voltage dependence, which you can reduce again by using even larger devices.

Again, don't get carried away by the impressive performance shown with a single sweep, which assumes perfect matching. A Monte Carlo run will show you the true behavior.

For CMOS Current mirrors there are three more sophisticated schemes. Figure 3-23 is the one you frequently see in articles. M1 is a thin device, producing a bias voltage about 100 to 200mV higher than the gate voltages of M3 and M5. Since the gates of M2 and M4 are connected to this point, these two devices act as cascodes, i.e. they shield the lower two devices from voltage changes.

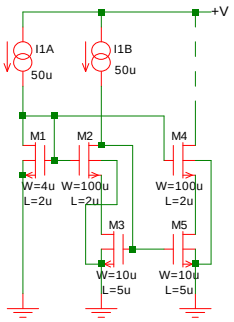


Fig. 3-23: Widely used mirror.

With the dimensions shown, the circuit in figure 3-24 has exactly the same performance with fewer devices and less current consumption; here the lower devices are dimensioned so that the gate voltage (at 50uA) has the required value for cascode biasing. The upper devices are then made wide enough to leave a comfortable margin between their source potentials and the "on" voltage of the lower devices (i.e. the voltage drop caused by channel resistance). This is a prime example how much you can do by simply changing the channel length and width of a CMOS transistor.

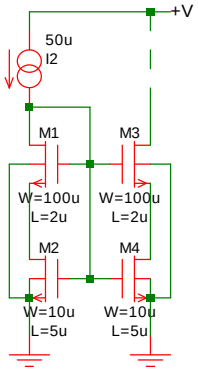


Fig. 3-24: Fewer devices, same performance.

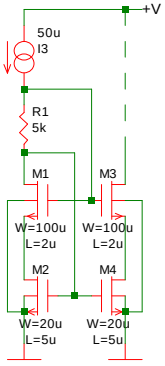


Fig. 3-25: Best Performance.

The circuit in figure 3-25 has the best performance. The cascode bias voltage is set not only by the device dimensions, but by the small (250mV) voltage drop across R1. Since the current itself is almost certainly determined by a resistor, R1 will track it. The flatter curve represents a higher output impedance (100MOhm for 3-25, 33MOhm for the others) which is important for high gain in active loads.

All the figures given here are for room temperature only. The threshold voltage, the resistor and, in bipolar designs, the VBE have temperature coefficients. Make sure you simulate your circuit over the entire temperature range.

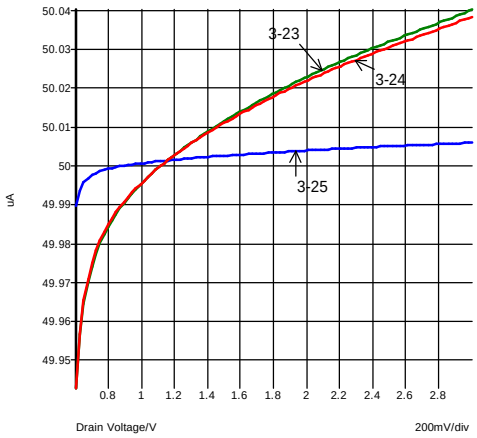


Fig. 3-26: Comparison of Performance.

4 The Royal Differential Pair

Open any analog IC and you will find a differential pair. Or, more likely, a half dozen. It has great advantages, even if amplifying a "difference" is not even a goal.

The reasoning is simple: Individual integrated components have large variations, but two (or more) of the same match very well. If you can take advantage of the matching, you get better performance.

It isn't always true, of course. Noise, for example can be smaller in a single-transistor stage and some of the most ingenious designs are remarkably free of the common differential stage. But let's look at this wondrous tool.

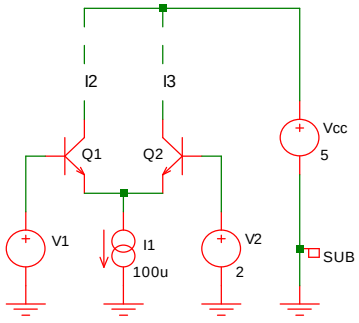


Fig. 4-1: In a differential pair a current is divided by two transistors.

Two transistors - here bipolar - share a common emitter current. If the voltages at their bases are equal and the two transistors match perfectly, I1 is split into two equal parts at the collectors, I2 and I3.

If we increase V1 (relative to V2), Q1 gets more of the current than Q2. If we decrease V1, the opposite is true.

But there are limitations and errors. First of all, the current division (or input voltage to output current relationship) is not linear. We are dealing with two base-emitter diodes here, fundamentally exponential devices. Not counting stray effects the emitter resistance is:

$$r_e = \frac{k * T}{q * Ie}$$

where k = Boltzman constant (1.38E-23 Joules/Kelvin)
T = the absolute temperature in Kelvin
q = the electron charge (1.6E-19 Coulombs)
Ie = the operating current through each emitter

This expression amounts to about 26 Ohms, at room temperature and with a current of 1mA. If Ie drops to 100uA, r_e becomes 260 Ohms.

Since it is very much a function of current, r_e is called the **dynamic emitter resistance**. The conversion from base voltage to collector current, the transconductance, is

$$g_m = \frac{1}{r_e + R_e}$$

where R_e is the ohmic resistance of the emitter, i.e. the resistance between the emitter contact and the emitter-base junction (usually a few Ohms).

As the current moves from one transistor to the other, *both* emitter resistances change and we get a rather non-linear behavior. Only a small portion of the curve in the middle, when the two current are equal or nearly equal, could be called linear, though in truth it too is not a straight line.

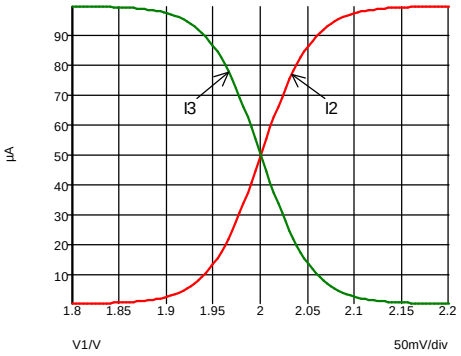


Fig. 4-2: The conversion from input voltage to current (the transconductance) is non-linear.....

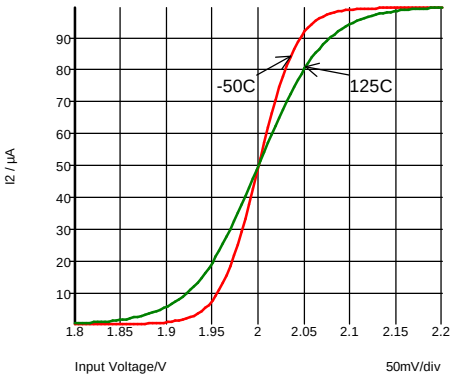


Fig. 4-3: ...and also temperature-dependent.

The other variable in the equation is temperature. The emitter resistance is proportional to absolute temperature, so at high temperature you get less gain or transconductance.

There are also three sources of error to be considered: 1. A small portion of the emitter current comes from the bases, not the collectors. With a minimum hFE of 100 this makes the sum of the collector current smaller than the emitter current by 1%, 2. Transistors only match well if the are treated identically. In this case that specifically means the collector voltages have to be the same. If one is higher than the other, its transistor will have a higher gain because of the Early effect; 3. Devices never match perfectly; there will be some differences in both VBE and hFE and thus some uncertainty in the voltage at which I2 and I3 are equal, showing up as an **offset voltage**.

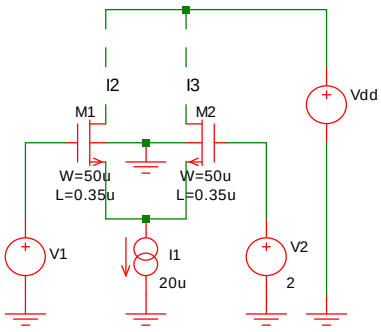


Fig. 4-4: MOS differential pair.

A differential pair using MOS transistors behaves almost identically, but *for entirely different reasons*. There is no dynamic emitter resistance; the gain is determined directly by the trans-conductance. This transconductance is also non-linear, increasing drastically with increasing gate voltage. Whereas in the bipolar transistor size is only of second-order importance, transconductance in an MOS transistor is directly proportional to gate width and decreases with increasing temperature.

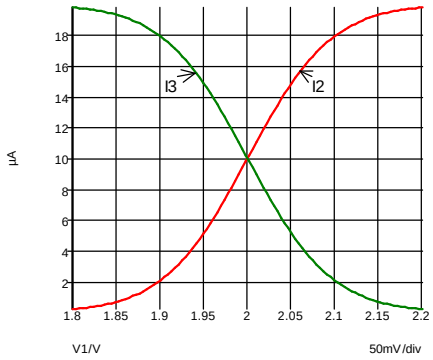


Fig. 4-5: A CMOS differential pair is also non-linear .....

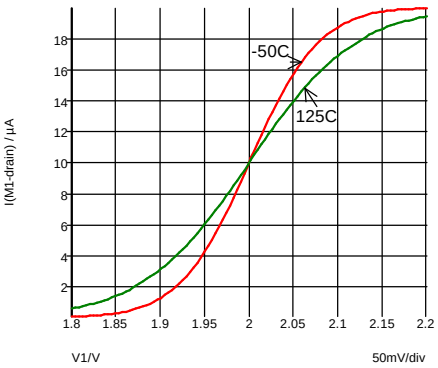


Fig. 4-6: ...and the transconductance also has a temperature coefficient.

Notably absent in the error sources is any kind of input current; I2 plus I3 are indeed equal to I1. There is, however, an offset voltage and, for equal sizes, MOS transistor have a larger offset voltage (i.e. mismatch) than bipolar ones (about 2:1, but this depends greatly on the process). Remember you can always improve matching (for any device) by increasing size (i.e. total area), preferably by using multiple small devices.

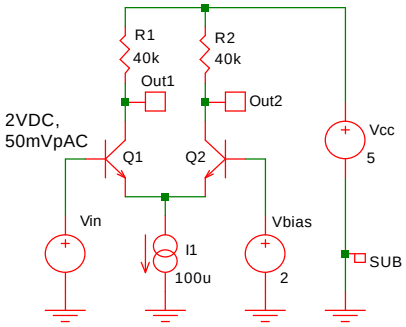


Fig. 4-7: A complete differential amplifier.

Let's get back to bipolar transistors and complete the differential stage. We can simply use the two collector currents to create voltages across two resistors. In this example the voltage at the base of Q2 is held constant - it is simply a DC bias

voltage large enough to overcome the base-emitter diode voltage (assuming that there is a single 5-Volt supply). Vin, going to the other base, carries the same DC bias level and has a 50mVp AC signal superimposed (i.e. it moves from 1.95 Volts to 2.05 Volts). The gain of this stage is determined by the ratio of the resistors to the dynamic emitter resistances. At 50uA (for each transistor) r_e is 520 Ohms (26Ohmsx1mA/50uA). The gain from the inputs (measured differentially, which in this case is simply Vin) to the outputs (again measured differentially, i.e. Out1 + Out2) is 80k/1.04k or 77. The gain to only one output is half of that.

It's not a great deal of gain and it cannot be made any larger by simply increasing the values of R1 and R2. There is a DC voltage drop of 2 Volts across them, if we were to double their values the transistors would saturate.

In reality the gain is always lower than obtained by this simple calculation, which does not take into account the ohmic (i.e. access) resistances in the emitters or the fact that a small percentage of the emitter current is lost to the base.

Even with only 50mVp input there is already significant distortion, about 5%. We can improve this by connecting resistors in series with each emitter. This makes the total emitter resistance more linear, which drops the distortion to less than 0.1% (with 50mVp input). But the gain has suffered badly: less than 16 with a differential output, less than 8 single-ended.

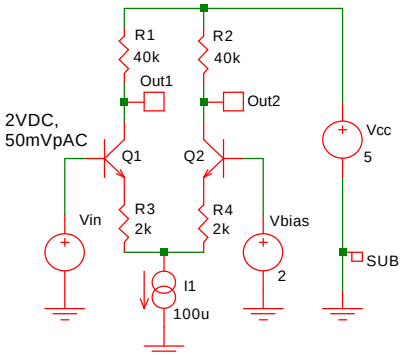


Fig. 4-8: Linearized differential amplifier

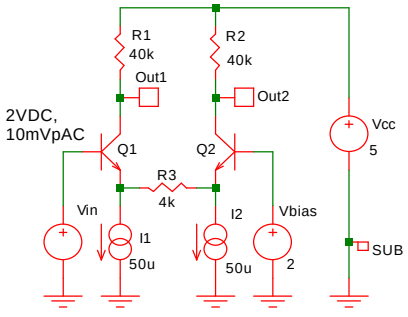


Fig. 4-9: A connection different from that of figure 4-8 but identical in performance.

There is a competing approach: two separate emitter current sources (or, more precisely, current sinks) of half the value and a single resistor of twice the value connected between the emitters. If you take a poll among analog IC designers about half of them will swear that one is better than the other. But in fact the two circuits are identical in performance.

To get more gain we need a better scheme for the output. In the vast majority of applications an amplifier needs only one output. Thus it is no loss

if we convert the differential signal to a single-ended one in the very first stage. And, with a current mirror, the benefits are immediate.

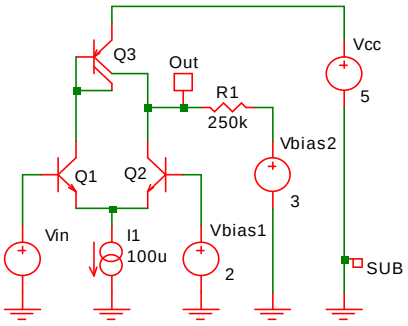


Fig. 4-10: Differential amplifier with an active load.

If one NPN collector current is mirrored by Q3 (here a split-collector lateral PNP) it opposes the current of the second collector. With no input signal (and perfect matching) the two are equal, they cancel each other. But with an input signal one increases, while the other one decreases (hopefully by the same amount), so that we only see their *difference* at the output and we can use a much larger output resistance. As drawn, the gain of this stage is 278. And since the gain is so large, we get quite a large output signal with only 1mVp at the input, which makes the distortion reasonably small. Q3 is called an **active load**.

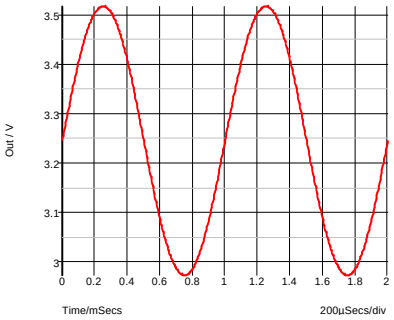


Fig. 4-11: Output signal with a 1mVp sine-wave input.

But there are two things wrong with the circuit in figure 4-10. First, if you were to specify a 250kOhm resistor in an IC you might be suspected of lunacy; its size would take up more space than all the other components together. Second, if you look at the output waveform closely, you notice that there is a DC current flowing through R1. At the right end we connect it to a 3-Volt bias point, but at the left end the center of the sine-wave is not 3 but 3.25 Volts, i.e. we have a built-in offset. There are two reasons for this: 1. the two collectors of Q3 are not at the same potential and 2. the collector current of Q1 has to supply the base current for Q3.

We need something like R1 to fix the DC potential at the output. The two opposing collectors are current sources/sinks. The smallest difference between the two would cause the output potential to move up so much that Q3 would saturate or down so much that Q2 would saturate.

In figure 4-12 all of these problems are fixed at once by adding a second stage. Q4 is the same size as Q3 and is operated at the same current. Now the collector voltages of both Q1/Q2 and Q3 are identical. Moreover, the collector current of Q2 has to supply the same amount of base current

(for Q4) as the collector of Q1 does. In other words, with no input signal the circuit is perfectly balanced; there is no built-in offset.

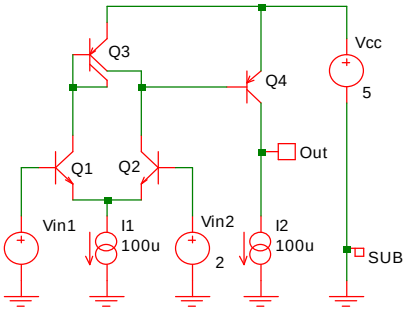


Fig. 4-12: High-gain, balanced differential amplifier.

But you need to be careful here. The gain of this circuit is no longer fixed by a resistor ratio, it is dependent on transistor parameters. If these two stages are made part of an operational amplifier, the feedback will take care of this. Or, if the circuit is used merely as a comparator, gain is of lesser importance than offset.

If you simulate the circuit as shown, you will get a different and rather odd output curve. Current sources in a simulator are ideal devices, they will do anything to supply the exact amount of current, which includes supplying their own voltage (if necessary thousands of volts). An actual current sink such as I2 will collapse near ground, but the ideal

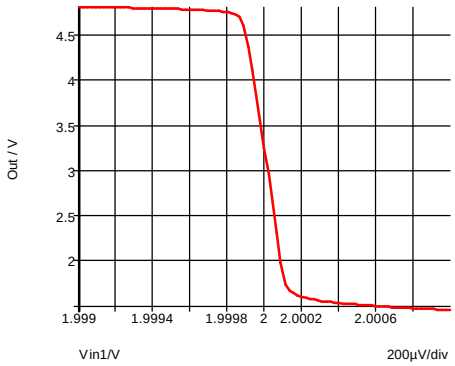


Fig. 4-13: Transfer curve for circuit in figure 4-12.

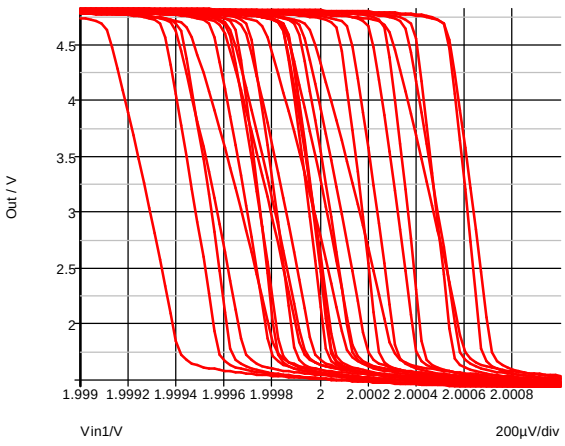


Fig. 4-14: Even a perfectly balanced differential amplifier (figure 4-12) has an offset voltage due to mismatch.

one keeps right on working down to a very large negative voltage. For this reason there was an additional device in the simulation diagram, a diode from Vin2 to Out which clamps the output swing at the low end.

The last circuit may be perfectly balanced but, as in any circuit, the matching of the devices is still subject to variation. If we run a Monte Carlo analysis we meet the real world: the random offset voltage.

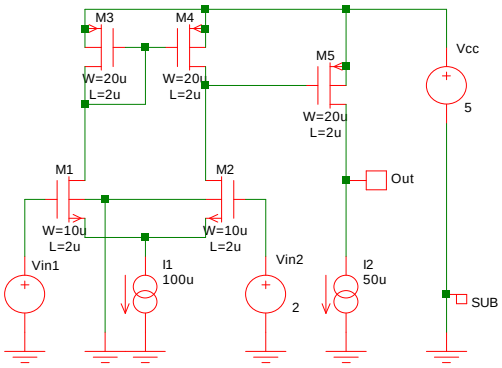


Fig. 4-15: Balanced CMOS differential amplifier.

Figure 4-15 shows the same design in CMOS. Note that M3, M4 and M5 are all the same size, thus balance is achieved with I2 having a magnitude of one-half I1. Here, of course, we are not concerned about cancellation of base currents but identical gate voltages are still important.

The random offset voltage is of greater concern in a CMOS design. MOS transistors match less well than bipolar ones. That has been true since the start of the IC industry. It is not that an MOS transistor is inherently inferior in this respect, but that matching, specifically the offset voltage is based on different process parameters. For the bipolar transistor matching is determined by the depths of diffusions, particularly the base and emitter. The dimension having the greatest influence on offset voltage in an MOS transistor is the gate insulator thickness. While control has steadily increased, the insulator thickness needed to be steadily decreased to get sufficient gain for the ever smaller devices. The gate insulator thickness has by far the smallest dimension in an IC and thus continues to create fluctuations in threshold voltage larger than a diffusion will cause in VBE.

Two more additions. The base current of a bipolar transistor is a disadvantage. If its operating current is say 25uA and the minimum hFE 100, the input draws (or supplies in the case of a PNP transistor) as much as 250nA. We can decrease this with a **Darlington** configuration.

Q3 and Q4 carry the base current of the differential pair. At their bases the input current is reduced by a factor of another hFE. Thus the input current is

$$Iin = I1/2(hFE)^2$$

or 2.5nA. There is a price, though:

1. The input voltages need to be higher by a VBE so that there is enough headroom for I1;

2. Q3 and Q4 run at very low current, thus their speed is bound to be

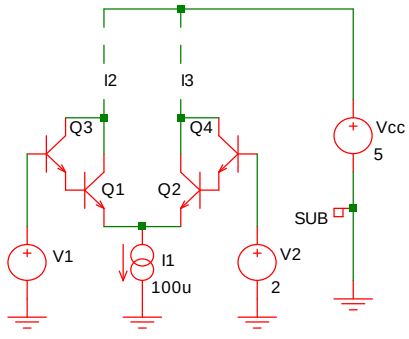


Fig. 4-16: NPN Darlington input stage.

rather slow; and

3. the leakage currents of Q3 and Q4 run into the bases of Q1 and Q2, showing up multiplied by the hFE of the latter two in I2 and I3. This is a danger at high temperatures (say above 90°C).

Switching time and leakage current can be reduced with small currents from the emitters of Q3 and Q4 to ground, in effect running the two transistors at a higher operating current. But of course you can't go too far in that direction: the input current increases again.

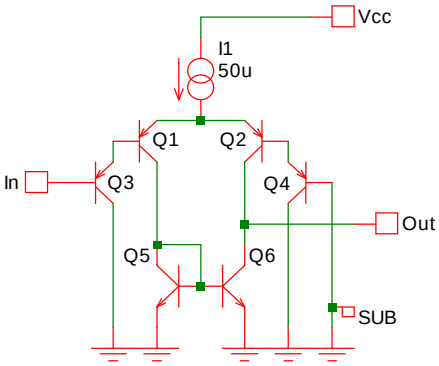


Fig. 4-17: A PNP Darlington input stage allows the input to move below ground.

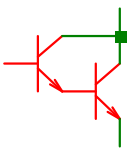
If you invert the circuit and use (lateral) PNP transistors at the input, you gain an advantage which is often useful: one input can be at ground. There is enough headroom for the current mirror (Q5 and Q6) even if the input is 200 or 300mV below ground. The limit here is one diode drop below ground, at which point the base of Q3 will forward-bias against the substrate. The same limitation in speed and upper temperature apply.

The Darlington Pair

Sidney Darlington was born in Pittsburgh, Pennsylvania, in1906 and joined Bell Laboratories in 1929, where he remained until his retirement 42 years later. He was a theorist who also liked to tinker with circuits.

In 1952 silicon transistors made at Bell Labs had low gain (5 to 15). Darlington checked out two of them (only a few were available) and experimented at home over a weekend. He found that by connecting the emitter of one to the base of the other, the gain would be the product of the two, 25 to 225, a much more useful range. He then suggested a method fabricating the pair out of a single block of silicon (with a common collector), thus coming very close to the idea of a monolithic integrated circuit. Bell Labs was issued a patent (2,663,806) in 1953.

Darlington died in 1997 at age 91.



5 Current Sources

Ever since the dawn of analog IC design (all the way back in 1962) a succession of very clever people have been trying to conjure up something that would produce an accurate current. The results have been uniformly dismal.

There happens to be a capable voltage source in ICs, the bandgap reference (which we shall get into next). So, to get a current, one would think, all one needs is an accurate resistor; after all I = V/R. But, unless you want to add a costly thin-film layer and laser trimming, there are no accurate resistors. What we get are resistors made from diffused or deposited silicon layers which vary in resistance from wafer to wafer and have a considerable temperature coefficient.

So, don't expect any precision here. At best, an integrated current source can provide a small current without the use of large-value resistors and make this current more or less independent of the applied voltages.

Current Sources with Bipolar Transistors

The first example uses a diode-connected transistor (Q3) as a reference voltage. A primary current flow through R2, Q2 and Q3. The base of Q1 is at two VBE (base-emitter or diode voltage), thus its emitter has a potential of one VBE. The current through Q1 is thus VBE/R1. If we let the voltage at the collector of Q1 (the destination of the current) drop below about 1 Volt, Q1 saturates and draws from the primary current. But above 1 Volt the current is very constant, changing less than 0.3%. This quality is best expressed as output

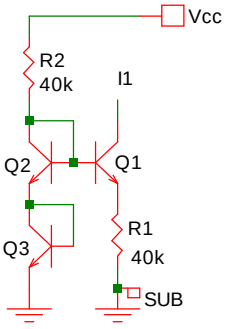


Fig. 5-1: Current source based on VBE (Q3)

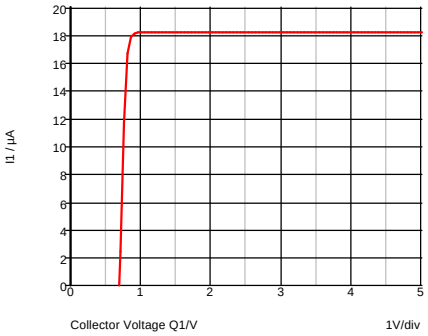


Fig. 5-2: I1 vs. output voltage.

impedance, i.e. the change in voltage (4 Volts) divided by the change in current (about 53nA). Thus the output impedance of this circuit is about 75MOhm. Not bad for using only 80kOhms in resistance.

A diode has a negative temperature coefficient and a (diffused) resistor a positive one. The two combine to give the current a strong negative temperature coefficient, a change of about -29% from 0 to 100^0C.

The VBE is of course no Zener diode, it varies a bit as the current changes, which makes the current dependent on the supply voltage (a +2% increase as the voltage moves from 4.5 to 5.5V).

And then there is the variation in production: ± 28%, mostly caused by the variation of R1. Changes with temperature and supply voltage must be added to this figure.

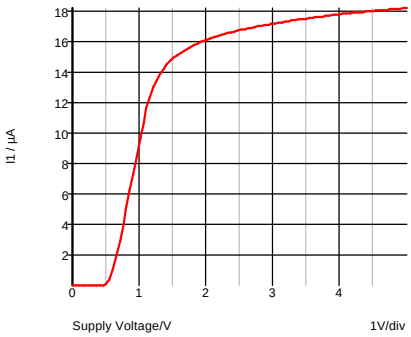


Fig. 5-3: I1 vs. supply voltage.

Also be aware that we are wasting some current: it takes 90uA through R2 to produce 20uA in Q1.

A word about the choices in the examples of this chapter:

- For bipolar circuits the use of (base) diffused resistors is assumed with an absolute variation of ± 25%. This is probably the largest variation you will encounter; CMOS foundries often guarantee smaller variations, especially for poly resistors.
- Each current source produces about 20uA, an arbitrary choice made to allow comparison.
- Strictly speaking these circuits are current sinks, not sources. To make a circuit in which the current is delivered from the positive supply, the design is turned upside-down, NPN transistors are made PNP and N-Channel ones P-Channel.
- Supply voltages are arbitrarily selected from 5, 3 and 1.8 Volts.
- As before, the fourth (bipolar) transistor terminal is hidden; all of these terminals are connected together to the most negative supply voltage with the symbol SUB. This avoids cluttering up the schematic. For MOS transistors the connection is left visible as a choice needs to be made for the P-Channel device.

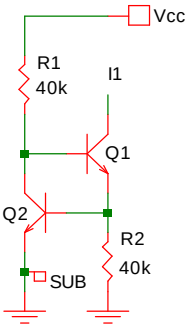


Fig. 5-4: Improved VBE current source.

On to the second example, a rare case where better performance is achieved with fewer devices. Through feedback Q2 regulates the current of Q1, holding I1 more constant. In this way the output impedance increases to 500MOhm, the Monte Carlo variation decreases to ± 26% and the change with supply voltage from 4.5 to 5.5V to +1.8%. But the voltage at the collector of Q1 still must not drop below about 1 Volt.

In both of these current sources the emitter of the output transistor is sitting on top of one VBE, about 0.65V at room temperature (and higher at low temperature). For low-voltage ICs we need a design in which this emitter is at or very near ground.

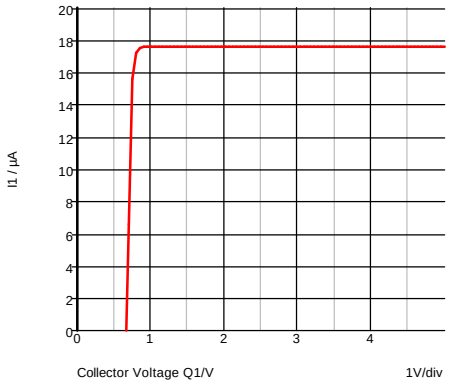


Fig. 5-5: I1 vs. output voltage of figure 5-4.

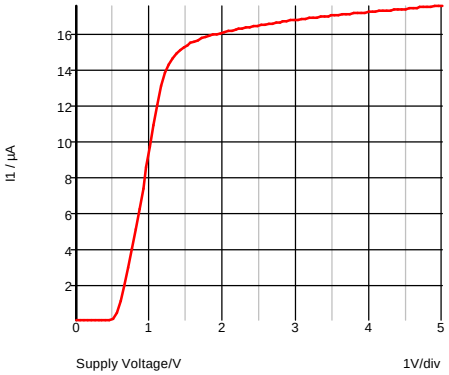


Fig. 5-6: I1 vs. supply voltage of figure 5-4

Figure 5-7 looks like a current mirror, but it isn't. There is a deliberate mismatch between the two transistors. They get the same voltage at their bases but, while Q2 has a straightforward base-emitter diode to ground, the path for Q1 consists of a *lower* diode voltage (because of the larger area using three emitters) and a resistor. The difference in voltage between the two diodes is:

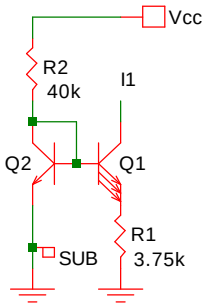


Fig. 5-7: Delta-VBE current source.

$$deltaVBE = \frac{k*T}{q} * \ln\left(\frac{A1*I2}{A2*I1}\right)$$

where k = Boltzman constant (1.38E-23 Joules/Kelvin)
T = the absolute temperature in Kelvin

q = the electron charge (1.6E-19 Coulombs)
A1 = emitter area of Q1
A2 = emitter area of Q2
I2 = current through Q2

k*T/q amounts to about 26mV at room temperature and I2 is about 110uA. Thus, with a desired I1 of 20uA, the voltage drop across R1 is 26mV*ln(16.5), i.e. I1 = 72.9mV/3.75k = 19.4uA. Note that *delta-VBE is independent of current, only the current ratio is important.*

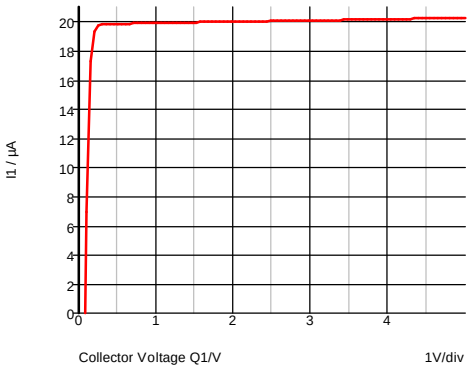


Fig. 5-8: I1 vs. output voltage of figure 5-7.

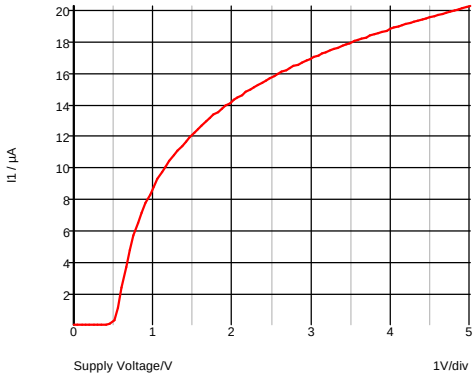


Fig. 5-9: I1 vs. supply voltage of figure 5-7.

The voltage at the collector of Q1 can now go lower, to the saturation voltage of the device plus the delta-VBE. There is little change in current as the voltage at the output is moved (amounting to an impedance of about 12MOhm) but dependence on supply voltage is quite large, a +6.5% change as Vcc moves from 4.5 to 5.5V.

Since delta-VBE is proportional to absolute temperature (**PTAT**), I1 has a marked positive temperature coefficient, moderated only slightly by the positive tempco of the resistors. Production variation (at a fixed voltage and temperature) is ± 26%, dominated by the variation of R1.

The performance can be improved slightly by using a larger device ratio. With Q1 having 10 emitters the output impedance increases to 15MOhms and the voltage dependence to 4.5% (4.5 to 5.5V).

The Quality of a Current Source

An ideal current source maintains the current level no matter what happens at its terminals, which results in an impedance that is infinite.

A practical current source can approach this over a limited voltage range, with an impedance of up to tens of Meg-Ohms (i.e. there is very little change in current as the voltage across the current source changes. But its absolute level is subject to (absolute) parameter variations, which are large in an integrated circuit.

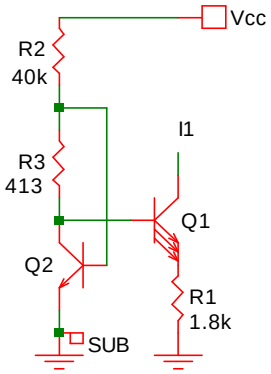


Fig. 5-10: R3 reduces supply-voltage dependence.

You can also improve this performance with a simple measure. Suppose you connect the base of Q1 not to the base of Q2, but to a point which *counteracts* a rising Vcc. By inserting a small amount of resistance in the collector path of Q2 we get a node whose voltage is fairly constant. The voltage at the base of Q2 still increases somewhat as Vcc is increased, causing its collector current to increase. But this makes the voltage drop across R3 increase and, with just the right value, the base voltage for Q1 changes little, at least over the critical range in supply voltage.

Note that, because of the lower base voltage, the value of R1 is lower for the same amount of current. The easiest approach to circuits like Figure 5-10 is simulation. Just try various values for R1 and R3 until you get the right current with minimal change. But , in the layout, make these resistors fairly wide; you are counting on matching.

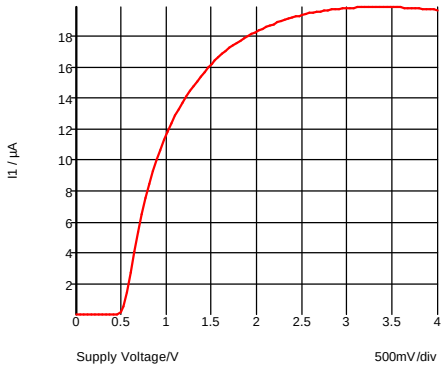


Fig. 5-11: I1 vs. supply voltage with R3 optimized for the range 3 to 3.6 Volts.

The change in I1 is now a mere ±0.5% with Vcc varying between 3 and 3.6V (and even lower with a 4.5 to 5.5V range).

The temperature coefficient for this circuit is somewhat larger: a +31% change from 0 to 100°C and the output impedance drops to about 7MOhms. Production variation is unchanged.

We are about to take a rather daring step. The primary current in the previous circuits is a nuisance; it wastes power and takes up considerable resistance. Why not replace it with a current source derived from the current the circuit generates?

There is one flaw in this argument: the current must *exist* first. There are two possible modes, one in which the current levels are as intended and one where there are no currents at all. In other words, there must be a current in Q2, which can be mirrored and fed back to Q1 and the base of Q2 so Q2 can have a current, etc.

The usual solution is to employ a start-up circuit, designed to bring Q2 to a level sufficient to sustain the loop. The start-up circuit then shuts down and has no further influence.

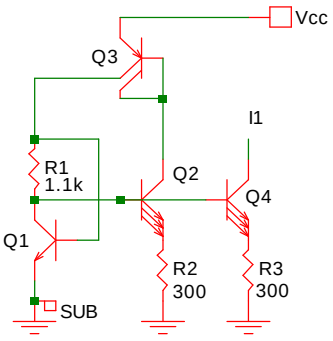


Fig. 5-12: Self-starting current source without large-value resistors.

But there is another way: leakage currents. Q2 has two leakage currents, from collector to substrate and from collector to base. These currents may be small (pA), but they are mirrored by Q3 and fed back into the base of Q2, where they are amplified. And so it goes around the loop, eventually reaching microamperes.

Two factors must be understood here. First, we are not talking about a leakage current caused by dirt. The very small reverse-junction currents measured in today's IC devices are fundamental phenomena and have nothing to do with cleanliness. Second, the design and the process must allow these small currents to grow. If, for example, there is a path from the base of Q3 to Vcc or the bases of either Q1 or Q2 to ground which can shunt leakage (provided say by a very large, reverse-biased junction), the scheme won't work. If your models are accurate, trust the simulation. Use a Monte Carlo analysis to see if the circuit starts up every time and do this at temperature extremes where leakage currents are either at their lowest or highest.

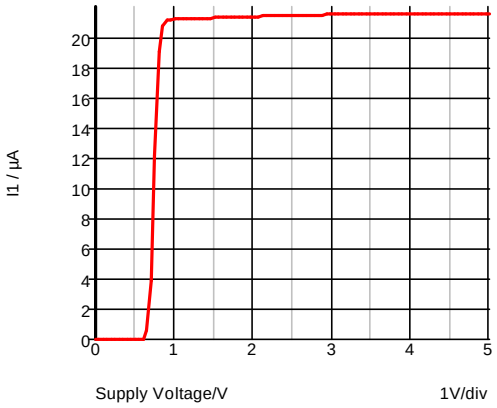


Fig. 5-13: I1 vs. supply voltage for figure 5-12.

R1 has been added to counteract the remaining dependence on Vcc (caused by the Early effect in Q2 and Q3). With that we get a change of +0.4% as Vcc is increased from 4.5 to 5.5V. The circuit can have a supply voltage as low as 1 Volt and the voltage at the output can be as low as 0.3V.

Lastly, the Erdi current source, a very clever design with an astonishing performance. We start with an auxiliary current, Iaux. And before you even have a chance to sneer at the fact that a current source is

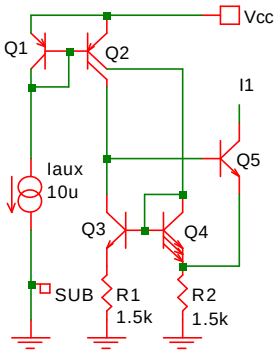


Fig. 5-14: Erdi current source.

used to make a current source, let me point out that the accuracy of this current source is of no great importance. A bulk (epi) pinch resistor will do or any of the lesser current sources discussed above.

Iaux is mirrored and split into two equal parts by Q1 and Q2; thus the operating currents for Q3 and Q4 are equal. Q4, however has 3 emitters, Q3 only one, thus there is a difference of about 29mV (at room temperature). Unbalanced, the collector voltage of Q3 rises until Q5 supplies enough current to make up the difference. This current amounts to (delta-VBE)/R2.

Vcc can be as low as 1 Volt (or as high as breakdown voltages allow). Moving Vcc 20% changes I1 by 0.08%. The output impedance is 50MOhm. Temperature is strongly positive, a +25% change from 0 to 100°C. And the Monte Carlo variation is roughly that of R2, here assumed to be ± 25%.

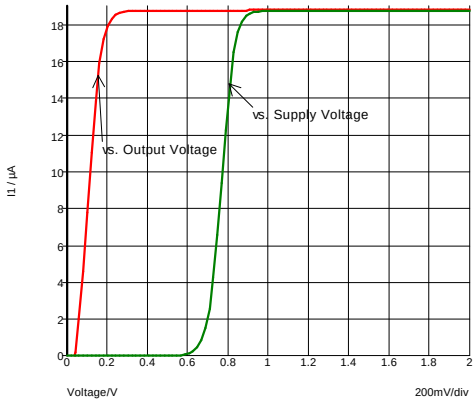


Fig. 5-15: Performance of the Erdi current source.

CMOS Current Sources

None of the bipolar schemes work well for CMOS devices. They are based on VBE or delta-VBE , for which there are no equivalents. Trying to use circuits such as Figures 5-10 or 5-14 with CMOS width-ratios leads to inferior circuits. Also, due to the square-law behavior of the gate voltage, the variations are roughly double those of bipolar designs.

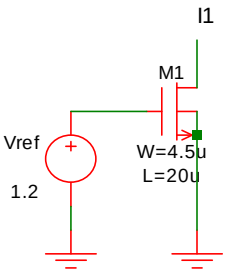


Fig. 5-16: MOS transistor as a current source.

Fortunately the CMOS transistor *is* a current source. If we simply apply a constant voltage to the gate (such as a reference voltage) we can tailor the width and length of the device to give us a certain current.

Using a rather exaggerated length, we can minimize the channel-shortening effect. In the example

here the output current varies little with the applied voltage, amounting to an impedance of 38MOhm. But the variation (due to the uncertainty in the threshold voltage) is large: ± 39%. Add to this the change with temperature (0 to 100°C, -23%) and a variation in the reference voltage (± 3% causes a change of 10% in I1).

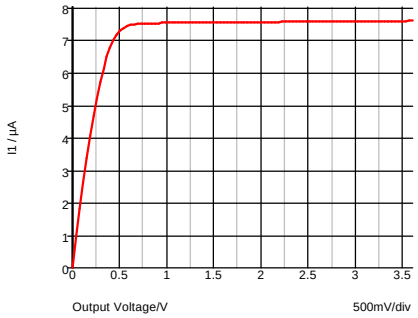


Fig. 5-17: I1 vs. voltage for fig. 5-16.

The Ideal Current Source

Sometimes a compromise is the best solution. If we allow just one component to be external to the IC and provide a pin for it, the performance of a current source improves dramatically. All other currents within the IC can then be derived from it with current mirrors and are thus inherently accurate.

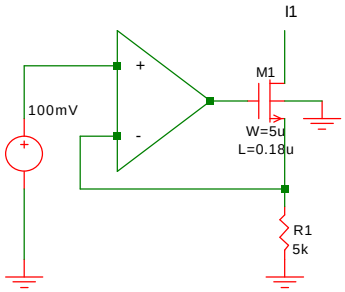


Fig. 5-18: Accurate current source with an op-amp and external resistor.

In the last circuit an op-amp compares the voltage across an external resistor with a low internal reference (divided down, for example, from a bandgap reference) and drives the gate (or base) of an output device. Assuming no trimming, a 1% tolerance for the external resistor, 3% for the reference voltage and 2mV offset uncertainty for the op-amp, I1 will be within 6% at any voltage and any temperature.

6 Time Out: Analog Measures

dB

Analog scales tend to be very large. As an example, the hearing threshold of a young adult is 20 micro-Pascal; the maximum level without damaging the ear can be more than 20 Pascal, a ratio of one to 1 million. Particularly because of the widely varying sound levels a need for a logarithmic measure appeared early on in electronics. There are two of them: The Neper is based on the natural logarithm and named after John Napier, a 16th century Scottish mathematician who came up with the logarithmic table (and whose name was most likely spelled Neper in his time).

In the 1920's a measure based on logarithm with base 10 began to be used at Bell Laboratories. At first it was called the "transmission unit", then re-christened the Bel, after Alexander Graham Bell. The idea is simple, a Bel is the logarithm of two power levels:

$$Bel = \log \frac{P1}{P2}$$

But the Bel turned out to be a bit coarse; one-tenth of that suited the Bell Labs people better, hence the decibel, or dB:

$$dB = 10 * \log \frac{P1}{P2}$$

This is the ratio of two power levels. Since power is related to the square of the voltage (or current), we get:

$$dB = 20 * \log \frac{V1}{V2}$$

Neper has more or less disappeared as a measure, dB proved to be more convenient. But, when using dB, always keep in mind there is a 2:1

difference between the ratios of power and those of voltages or currents (or pressure).

Even though a logarithmic ratio is very convenient, it is helpful to picture the (voltage) actual ratios:

-60dB	1/1000
-40dB	1/100
-20dB	1/10
-6dB	0.5 (exactly: 0.5012)
-3dB	0.707
0dB	1
20dB	10
40dB	100
60dB	1000

Fundamentally dB is relative, expressing only a ratio. But there are many modifications in which one of the two levels are absolute, among them:

dB(SPL)	Sound pressure level, where the hearing threshold (0dB) is based on 20uPascal.
dBm	Power ratio where 0dB = 1mW. Originally this was based on an impedance of 600 Ohms (that of a telephone line), but now is used with any impedance (which is fine as long as we calculate power ratios, not voltage ratios).
dBuV	Voltage ratio relative to 1uV.

RMS

RMS calculation was introduced by Charles Steinmetz more than 100 years ago. Steinmetz grew up in Breslau, Germany (now Poland), but shortly before he got his Ph.D. in mathematics and physics he had to flee to Switzerland because of his socialist activities. From there he emigrated to the U.S. and found a job as an assistant draftsman in Yonkers. The company fabricated hat-making machinery, but soon expanded into electrical motors. The year was 1889.

Steinmetz was a small man with a hunchback and one leg shorter than the other, a deformity he inherited from his father. Though he made the proverbial bad first impression, the people around him soon were in awe of his razor-sharp mind. No surprise then that the overqualified draftsman was

at the forefront in AC engineering within four years. Through mergers and acquisitions he found himself to be working for General Electric as head of the calculating department in Schenectady and teaching at Union College. He led a bohemian life; afraid to marry because of his inherited deformity, he shared his house with an entire family and kept several crows, raccoons, eagles, owls, squirrels, dogs and alligators. He resumed his socialist activities, expounding his ideas in a book; he was against competition and advocated an industrial reorganization by the government. It is remarkable that he got along very well with his bosses at GE. In all respects he was a delightful man who seemed to have a very happy life. Almost single-handedly he moved electrical engineering from a craft to a profession.

Steinmetz found that few "electricians" used mathematics, his specialty. The first curriculum in electrical engineering had started at MIT in 1882 and very few people understood AC. George Prescott lamented in 1888: "It is a well-known fact that alternating currents do not follow Ohm's law, and nobody knows what law they follow."

For example, there were two things wrong with the "average" value meters of the time displayed: in the first place the true average of an AC waveform should have been zero; in the second place the product of the average current and voltage gave the wrong answer for power. Also, a phase shift between voltage and current left almost everyone perplexed.

In 1893 Steinmetz presented his first paper on the use of complex numbers in electrical engineering. It was heavy going, delivered in a thick German accent. But he kept at it in paper after paper and then a massive, three-volume text book. By 1901 he had it down pat and published a textbook that was finally easy to understand.

What he said was this: In order to calculate the power correctly, you need to square the voltage (or current), calculate its mean (average) and apply the square root. Hence RMS - root-mean-square. Or a bit more detailed: You divide the waveform into equal segments over one period, square each segment, add up the squared values, calculate the average and take the square root of that.

The power, by the way, is the average power. There is no such thing as RMS power, only RMS voltage and RMS current.

For a pure sine-wave this works out as:

$$Vrms = \frac{Vpeak}{\sqrt{2}} = 0.707 * Vpeak$$

Here is a simple illustration of RMS calculation: Four time segments of 100usec each. First a voltage is at 5 Volts, then a zero, then at 2 Volts and finally at zero again.

$$Vrms = \sqrt{\frac{5^2 + 0^2 + 2^2 + 0^2}{4}} = 2.69V$$

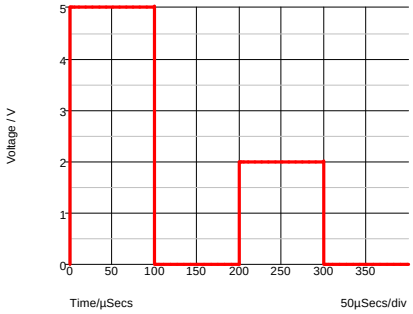


Fig. 6-1: Arbitrary waveform for RMS calculation.

But the RMS calculation has some limitations: it doesn't work with non-linear elements. In Steinmetz's time there were no transistors, not even vacuum tubes. There were only linear elements (save perhaps for the occasionally saturating transformer), so he didn't consider what would happen if the impedance changes while you are measuring RMS voltage and current.

Take the case of a transistor stage, either linear or switching. You want to determine its power dissipation, so you measure the current through it and the voltage across it. But the impedance of the transistor constantly changes and Ohm's law doesn't hold. The product of RMS voltage and RMS current gives an absurdly wrong result for power. The only way you can determine the power is to integrate the instantaneous values of voltage times current. In a simulation, Spice does this very well.

But measurements aren't nearly so easy: "True RMS" instruments do indeed have a circuit element which measures RMS rather than average. But the inputs to almost all of them are capacitively coupled. If the waveform you are measuring has a DC component, it is ignored and the result reflect on the AC portion of it.

Noise

Imagine a current flowing through a wire connected between a negative terminal on the left and a positive terminal on the right. Through the wire are flowing millions of electrons from left to right.

Each electron carries a charge of 1.6e-19 Coulombs. Let's say we observe a current of 1uA, thus 7.8e12 electrons pass every second. If the interval between electron were the same, we would then see a ripple at 7800Ghz, like the teeth of a saw-blade moving at high speed.

But in a diode or a bipolar transistor, we see quite a different effect. Here the current is initiated by electrons and holes moving across a barrier and this movement is anything but smooth. Within any given time interval one electron or hole may cross the barrier, or 5, or 100, or none. The variation is so great that there is no discernable peak at 7800Ghz, or any other frequency. In fact, because of the very large number of electrons, this **white noise** or **shot noise** is so well distributed over the frequency range that it has a constant level over the entire spectrum:

$$Inoise(rms) = \sqrt{2 \cdot q \cdot I \cdot B}$$

where q = electron charge (1.6e-19 Coulombs)
I = dc current
B = bandwidth in Hz

There are two things you should notice here. First: *Noise increases as the square-root of bandwidth*. Second: *When the current is decreased, noise becomes a larger fraction of it.*

Let's illustrate the second part. With a bandwidth of 10kHz, 1mA dc produces 1.8nA(rms) of noise. That amounts to 0.00017% or -115dB. With 1uA of current and the same bandwidth the noise is 56pA(rms), i.e. 0.0056% or -85dB. At 1nA we get 1.8pA of noise, which amounts to 0.18% or -54dB. All of which shows that it is harder to design a low-noise circuit at low current levels.

There is also noise when no current flows at all. By the energy imparted by temperature, some electrons will suddenly leave an orbit and jump to another. The higher the temperature, the larger this irregularity becomes. Thus a resistor, doing nothing but lying on a bench actually has a noise voltage at its terminals:

$$Vnoise(rms) = \sqrt{4 \cdot k \cdot T \cdot R \cdot B}$$

where k = Boltzmann constant (1.38e-23 Joules/T)

Thus a 1MOhm resistor always has a noise voltage of 13uVrms at room temperature, if measured over a bandwidth of 10kHz. This is called the **Johnson Noise**.

Whenever you mention a noise voltage or current, you also have to state the bandwidth. To avoid this, noise voltage is often expressed in **nV/rtHz** (nanovolts per root-Hertz). To get the real noise voltage you simply multiply this value with the square-root of the bandwidth.

These two noise sources are fundamental, present in any current or resistor. But there is another one, not fundamental exactly, but always present. It is called **1/f noise** or **flicker noise**.

Flicker noise is worst in an MOS transistor, which is a major reason why bipolar transistors are preferred in analog design. The silicon-oxide interface is capable of holding some electrons for a considerable period (seconds) and then releasing them in bunches. This increases noise at low frequencies far above the white-noise level; at 1Hz the noise level (in nV/rtHz) can be two orders of magnitude higher than at 1MHz.

Flicker noise is also present in bipolar devices, but to a lesser extent.

Fourier Analysis, Distortion

Jean Fourier was a mathematician who was active in the French revolution. He was arrested twice in the fight between the various factions but was spared the guillotine. In 1798 he joined Napoleon's army in the invasion of Egypt and then was appointed prefect in Grenoble; in 1809 Napoleon made him a baron.

In between his political and administrative duties he found time to not only publish a massive work on ancient Egypt but do mathematical research. He analyzed the flow of heat in mathematical terms, coming up with a novel expansion of functions as trigonometrical series. His memoir "On the Propagation of Heat in Solid Bodies" was read to the Paris Institute in 1807. His method, now called the Fourier series, was criticized by the leading French mathematicians and was not published until 1822 (and not translated into English until 54 years later). It turned out to have applications in a wide range of areas, including now electronics.

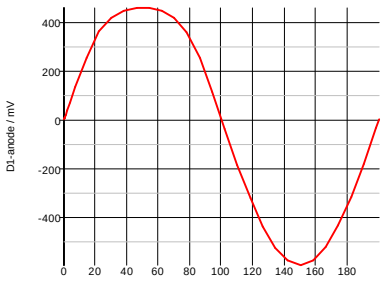


Fig. 6-2: Distorted sine-wave.

When a sine-wave is distorted other, higher frequencies are created which can be extracted in a Fourier series. There is an algorithm called "**Fast Fourier Transform**", or **FFT**, which does this. FFT uses an algorithm which allows fewer computations compared to the original discrete Fourier transform.

In our example here the positive half of a 5kHz sine-wave has been compressed. Converted into a frequency spectrum with the help of a Fourier transform we see the fundamental frequency of 5kHz and a series of

harmonics, multiples of the fundamental at 10kHz (second harmonic), 15kHz (third harmonic), 20kHz (fourth harmonic), 25kHz (fifth harmonic) etc., with gradually decreasing amplitudes. The square-root of the sum of the squares of all harmonics divided by the amplitude of the fundamental is the amount of distortion. You can usually disregard harmonics after the fourth or fifth, since their amplitudes become very small:

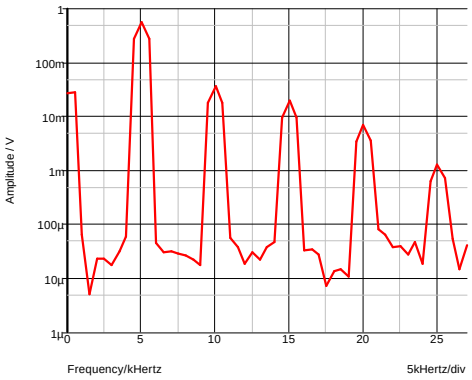


Fig. 6-3: Fast Fourier transform with low resolution.

$$Harmonics = \sqrt{36mV^2 + 21mV^2 + 8.2mV^2 + 0.9mV^2} = 42.5mV$$

$$Fundamental = 550mV$$

$$Distortion = \frac{42.4}{550} = 0.077 = 7.7\%$$

The peak at zero frequency shows the DC level, i.e. the asymmetry caused by the clipping.

Before you run a fast Fourier transform in Spice you need to choose two settings: how many samples should be taken over one period of the waveform and how many periods should be analyzed. In the example of Figure 6-3 there are in fact too few samples and periods, resulting in broad peaks.

As a general rule start with 25 samples and 50 periods. The first is determined by "maximum time-step" and "maximum print step" (set at 8usec in figure 6-4 for a 5kHz driving frequency) and the second by the total time in the transient analysis (1msec for 50 periods at 5kHz). You get the best results if both the number of samples per period and the number of periods are integers.

The fast Fourier transform has some flaws and limitations. For example, figure 6-4 shows peaks in between the harmonics, which, in

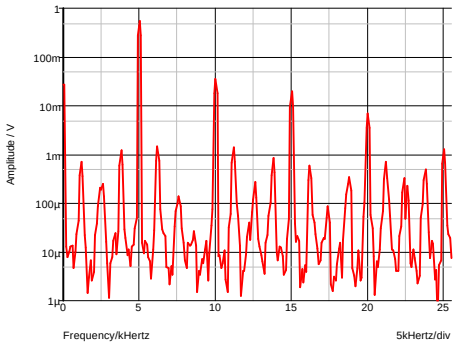


Fig. 6-4: Fast Fourier transform showing false peaks because of still insufficient resolution.

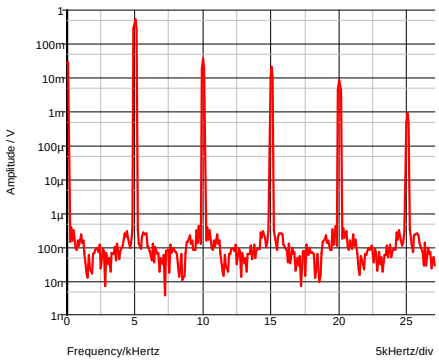


Fig. 6-5: Continuous Fourier transform with high resolution.

reality are not there. A superior method is the **Continuous Fourier Transform**, available in some analysis programs and shown in figure 6-5.

When you have a waveform which is symmetrical but not a sine-

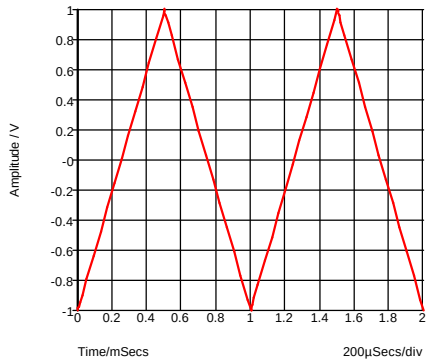


Fig. 6-6: Triangle wave.

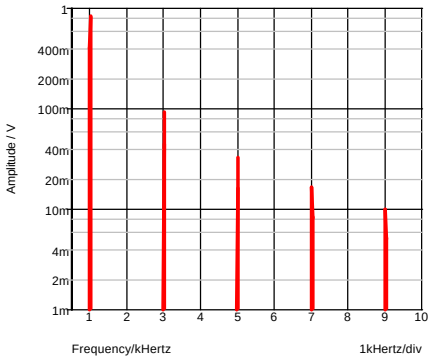


Fig. 6-7: Fourier analysis of a triangle.

wave, you get only odd harmonics (i.e. the third, fifth, seventh etc.). Shown here is the example of a triangular wave. Total distortion (i.e. deviation from a sine-wave) is 12%.

When you have non-linearity in a circuit and two frequencies are present, you also get **intermodulation distortion**, i.e. not only harmonics are created, but differences as well.

Frequency Compensation

Feedback is a wonderful thing. We take the inverted output signal, subtract the input signal from it and the amplifier will automatically correct and difference between them. If we only feed back a fraction of the output signal, the amplifier will automatically adjust its gain to one over that fraction.

Using a single frequency (any frequency), inverting a signal (negative feedback) is the same as a 180 degree phase-shift. And here comes the problem: Each device in the amplifier has a little bit of delay. At low frequency this has little effect, but as we go higher and higher in frequency the delay becomes more and more noticeable. At some high frequency the delay amounts to half a period of the signal and thus causes a phase-shift of 180 degrees. What started out as negative feedback now becomes positive feedback and the whole thing oscillates.

Frequency compensation is a design method which avoids this. The principle is very simple: deliberately slow down one device so that it is much slower than all others, i.e. it dominates the frequency response so that the delay in all other devices is no longer important.

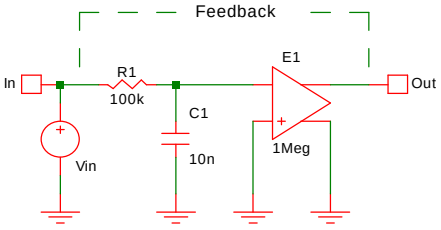


Fig. 6-8: Abstract circuit to illustrate phase-shift in a feedback amplifier.

This is illustrated with a very simple simulation. E1 is a "voltage-controlled voltage source" and acts like an ideal op-amp with a gain of 1 million (120dB), has no delay and the input and output terminals are free-floating (but are referenced here to ground). R1 and C1 cause the single delay (i.e. phase-shift).

Due to the RC network the amplitude at the output starts decreasing at about 100Hz. At this point the phase of the signal at the output is considerably less than 180 degrees, but as we go higher in frequency the phase never goes below 90 degrees. Thus the signal being fed back to the input cannot reach a phase-shift of zero degrees, the condition for

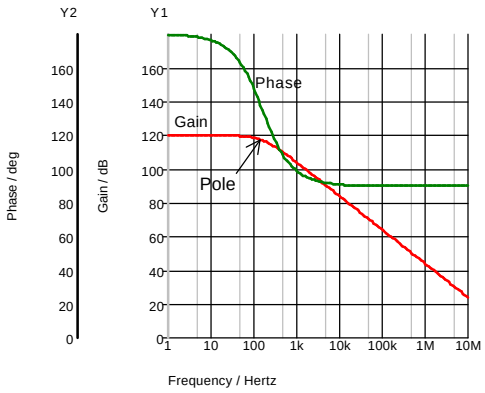


Fig. 6-9: Single-pole response. The phase never goes below 90 degrees.

oscillation.

The point at which the phase has turned by 45 degrees is called a pole. At frequencies somewhat higher than the pole the amplitude drops by 6dB per octave (doubling of frequency) or 20dB per decade.

Now let's look at the same simulation with another RC network added at the output, with a pole at a much higher frequency (C = 100pF). We now have two poles; you can just barely see the second pole (at about 10kHz) by the change in the steepness of the gain curve. The maximum phase-shift now is 180 degrees. The point of interest is the frequency at which the gain moves through zero dB (i.e. a gain of 1). If the gain is less than 1 an oscillation cannot sustain itself. While the phase at this point only approaches zero degrees, the margin is far too close for comfort.

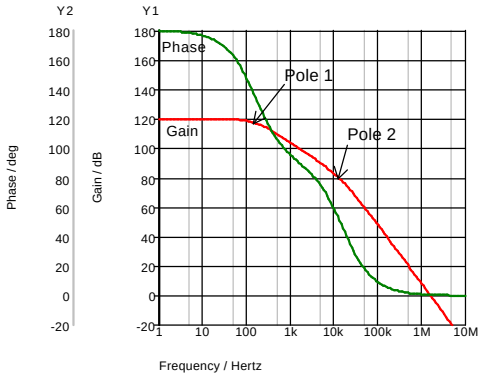


Fig. 6-10: Two poles in a feedback path approach zero degrees phase-shift.

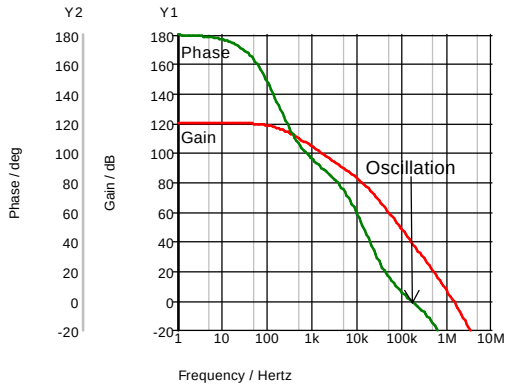


Fig. 6-11: Three poles in a feedback path. Phase-shift goes through zero degrees and oscillation takes place.

With three poles we are clearly out of luck. The phase now reaches zero degrees a decade before the gain drops below 0 dB. An amplifier which has these three poles will oscillate, in fact we can tell with certainty that it will oscillate at 200kHz.

There are now three remedies: 1. we can lower the gain until it drops below 0dB before the phase reaches zero degrees; 2. we can insert a new pole at a frequency so low that it dominates the others and 3. you can introduce a zero.

To illustrate the effect of a zero, we use another artificial circuit. R1/C1, R2/C2 and R3/C3 provide the three poles, delaying the phase of the signal, each by the same amount as in figure 6-11. R4, together with C2 provides the zero, it advances the phase rather than retarding it. The

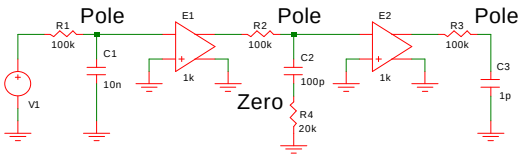


Fig. 6-12: Three poles and a zero.

(5MHz) the gain drops below 0dB but the phase is still positive, about 15 degrees, called the **phase margin**. Theoretically a feedback circuit with this behavior will not oscillate, though the phase margin is rather low. Since gain and time constants are subject to variation in an IC, it should be at least 60 degrees.

Now let's look at a real design, a simple, bipolar op-amp; this rather outdated

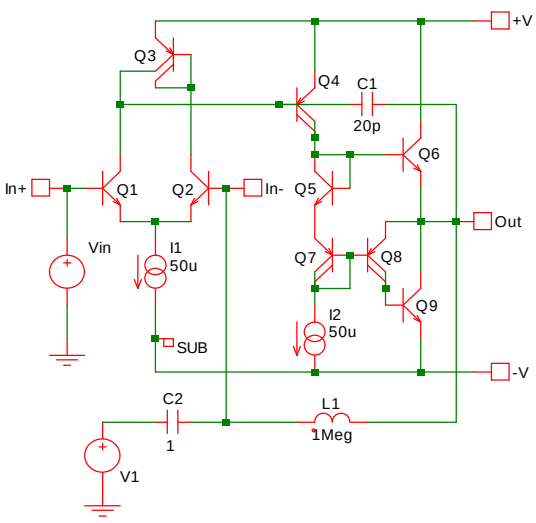


Fig. 6-14: Measuring gain and phase in a feedback loop.

frequency (i.e. the value of R4) is selected to result in a frequency where it is most effective.

At about 30kHz R4/C2 start turning back the phase, so that at the critical frequency

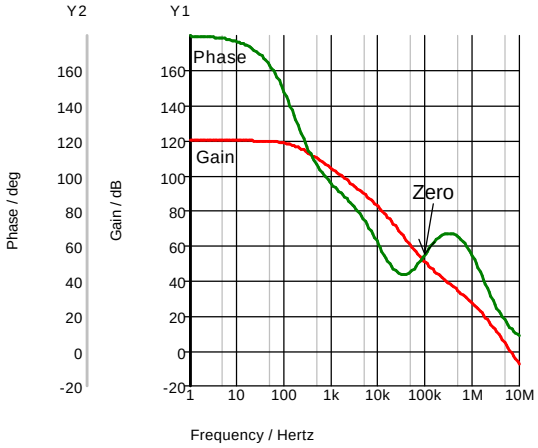


Fig. 6-13: A zero retards the phase-shift.

circuit was chosen because it uses the slow lateral PNP transistors, which aggravate the phase-shift problem beautifully.

The circuit uses the classical 3-stage design for op-amps (more of this in chapter 8): an input stage which converts the differential input signal to a single-ended one and has gain, a second stage (Q4) which provides more gain, and an output stage which has no (voltage) gain but provides a reasonably high output current. Since high-current PNP transistors are often not available in an IC, the lower portion of the output stage uses a compound transistor;

from the second stage it looks like a PNP transistor, from the output like an NPN one (but the combined device is achingly slow).

Q5 and Q7 are diode-connected transistors to bias Q6 and Q8.

The amplifier is investigated as a buffer, i.e. with a gain of one, produced by connecting the output directly to the inverting input. Here, though, there is an inductor in the path, which blocks AC but lets DC through so that the circuit is properly biased. C2, a very large capacitor, couples an AC signal to the negative input. In this way the feedback loop is opened up and we can measure loop gain and phase. This can be done at any convenient point in the loop, but the output to input connection is clearly the most convenient. Note that L1 and C2 have impractically large values. This is of no great consequence since these components are not going to be part on the design; we want to make sure they don't influence the AC behavior of the circuit.

We feed the AC signal into the loop after the inductor and then measure the loop response before the inductor (at "Out").

First let's look at the loop without C1. The loop gain is about 92dB and the phase drops rather sharply, reaching zero degrees long before the gain reaches 0dB. (Gain and phase have identical scales for easier reading). In fact, when the phase reaches zero degrees, the gain is still about 42dB. Therefore this circuit is unstable, it will oscillate.

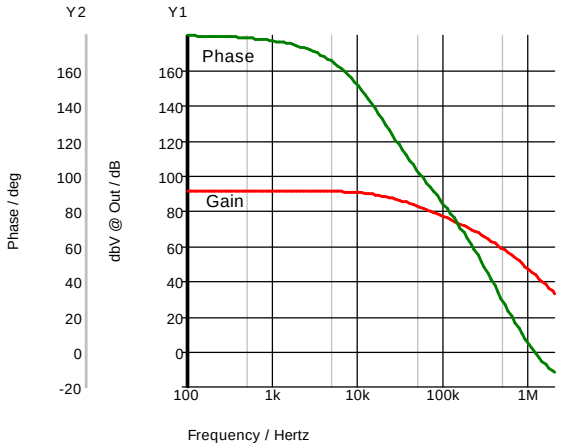


Fig. 6-15: Loop gain and phase of figure 6-14 without C1. The circuit would undoubtedly oscillate, the phase reaches zero degrees while there is still gain.

C1, the compensation capacitor, has been placed at the most strategic point in the circuit. There is considerable voltage gain between the base of Q4 and the output, which multiplies its apparent value (the Miller effect). Without this multiplication we would need a capacitor of about 2000pF, too large for an IC. It is also important that the capacitor feed back the AC signal from a reasonably low impedance (here the output) to a very high one (the current mirror and the base of Q4) so that we get nearly the full AC voltage swing at this point.

The result is self-evident. A new pole is created, about 100 times lower in frequency than the next higher one. This pole now dominates up to at least 10MHz and the phase is still 65 degrees away from zero when the gain drops below one. A stable circuit with an adequate safety margin.

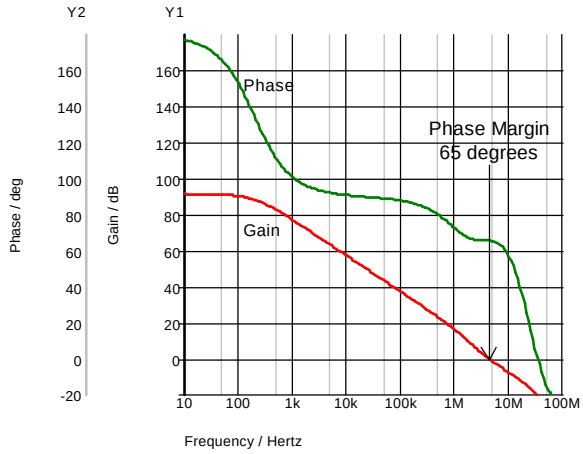


Fig. 6-16: With C1 the circuit of figure 6-14 has a phase margin of 65 degrees, i.e. the gain drops through 0dB safely before the phase reaches 0.

Of course there is a price to be paid for this stability: the gain of the op-amp may be more than 90dB at 10Hz, but it drops steadily as the operating frequency is increased. If we use this op-amp at 10kHz, we only have about 58dB of gain.

This analysis has assumed that the op-amp is going to be used with a gain of one. But if you are creating a design with a fixed gain, say 40dB, there is no reason why it should have to be stable at a gain of one. Which makes frequency compensation much less demanding. Just look at figure 6-15. Subtract 40dB from the gain curve (only the *excess* gain counts) and the amplifier is almost stable, i.e. a much smaller compensation capacitor is required.

The gain/phase analysis, as elegant and informative as it is, has a serious flaw: it shows performance only at one particular operating point (it is, after all, an AC analysis which does not disturb DC operating voltage and currents). A real-life signal will change the DC operating point and the loop gain and phase can change substantially.

Some simulators let you perform this AC analysis at different DC operating points, but there is an easier way , one that is a surefire test for stability. Get rid of the inductor and C2, close the feedback loop as intended in the application and apply a square-wave at the input. The square-wave should have fast edges (the default values in the simulator are adequate).

Then observe the output and watch for overshoot. For this circuit, with C1 at 20pF, there is a slight overshoot, one peak only. This circuit is very stable. (You can also see that the large compensation capacitor affects the slew-rate rather badly).

With the compensation capacitor reduced to 5pF there are three to four peaks, a damped oscillation. Up to four peaks are acceptable. If there are more, you are asking for trouble.

To make absolutely sure, do this with a brief (10 run) Monte Carlo Analysis at the temperature extremes and also for a rapidly varying load and supply voltage (less likely to cause instability, but it doesn't take much time to check). If there are never more than four peaks, you are safe.

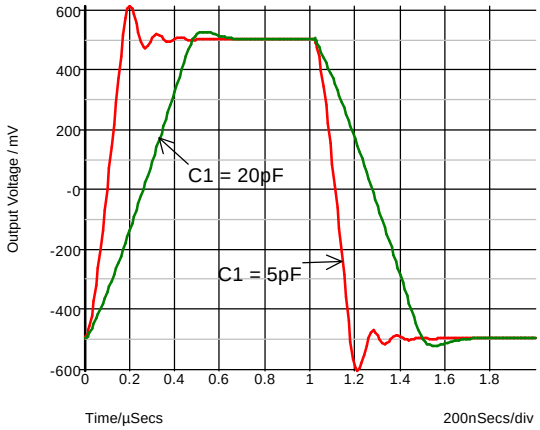


Fig. 6-17: To make sure a feedback circuit does not oscillate observe the pulse response. If there is ringing with fewer than 4 peaks, the circuit is stable.

A final small hint: in a gain/phase analysis simulators often get confused about the phase. You will see a plot which starts not at 180 degrees, but at -180. The two are in fact the same.

7 Bandgap References

In February of 1964 David Hilbiber of Fairchild Semiconductor presented a paper at the Solid State Circuits conference on "A New Semiconductor Voltage Standard". Zener diodes were still very poor and he was looking for something that drifted less over time.

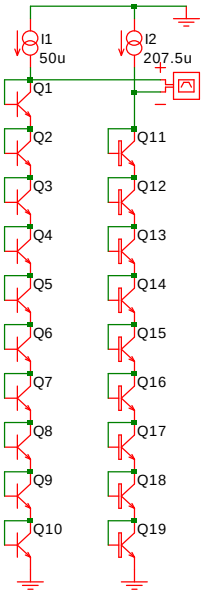


Fig. 7-1: The ancestor (1964).

It was already known that transistors with base and collector connected together made almost ideal diodes. Hilbiber took two of Fairchild's discrete transistors with greatly different forward voltages (which he attributed to different diffusion profiles) and made two strings with different numbers of transistors. He found a current level at which - over a narrow temperature range (± 2.5°C) - the voltage difference between the two strings changed little and amounted to 1.2567V. He attempted to find a relationship between this voltage and the bandgap potential of silicon at zero Kelvin, but found that it was primarily a function of the semiconductor material used in the two different transistors. He got what he was after, a much better long-term stability, and stopped at that.

Nothing happened for six years, when Bob Widlar put in the missing pieces. He recognized that the difference in diffusion profiles was only a secondary effect and the idea would work better if the two transistors were made by identical processes.

If you plot the diode voltage (VBE) over temperature you will notice that it points at the bandgap potential at absolute zero.

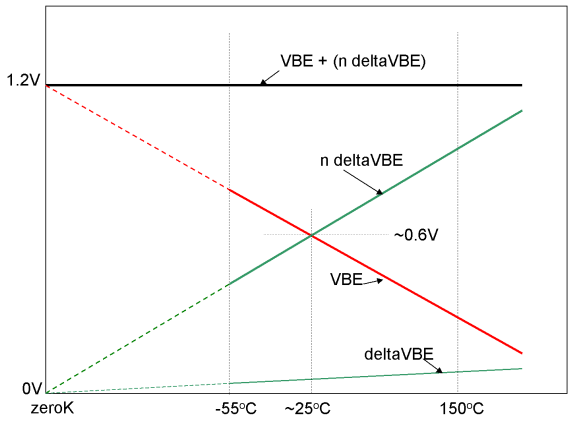


Fig. 7-2: The principle of a bandgap reference.

This is not strictly a straight line; it is slightly convex below about 150°C and concave above (it asymptotically approaches zero volts). The bandgap voltage at zero K, by the way, is strictly a theoretical concept; at that temperature there are no semiconductors, in fact electrons don't move at all.

Widlar found that an equal but opposite temperature coefficient can be created by running transistors at *different current densities*:

$$deltaVBE = \frac{k*T}{q} * \ln\left(\frac{A1*I2}{A2*I1}\right)$$

where A is the area (effective emitter area) of each transistor and I the current running through it. Here you have the choice of either using different emitter sizes, different current levels or both at the same time.

Delta-VBE is a true straight line, pointing to zero at zero K. But it is relatively small. kT/q amounts to about 26mV at room temperature, so an area (or current) ratio of 10 gives you a delta-VBE of about 60mV. As you can see from the diagram in figure 7-2 you need about 600mV at room temperature so it counteracts VBE.

But Widlar came up with a simple solution: multiply delta-VBE with a resistor ratio. R1 creates a current in Q1. Q2 has ten times the emitter area of Q1, so there is a delta-VBE between the two transistors of about 60mV (at room temperature). This delta-VBE shows up across R2. Ignoring a small error due to the base current, emitter and collector currents of Q2 are equal. Thus the voltage drop across R3 is delta-VBE multiplied by the ratio of R3/R2. Adding to this voltage the VBE of Q3 we get Vref.

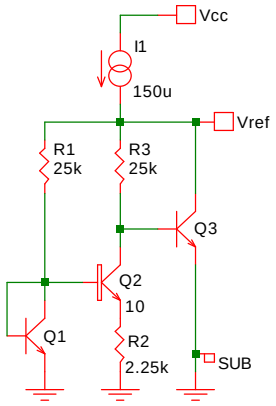


Fig. 7-3: Widlar's first bandgap reference.

The three transistor form a feedback loop (with limited gain, thus the internal capacitances are sufficient to keep it from oscillating), holding Vref at a constant level. If we increase the value of R3, Vref increases and the temperature coefficient becomes more positive. If we decrease R3, the opposite happens. In this way we can find the right value for R3 so that the negative temperature coefficient of the VBE is cancelled by the positive one of delta-VBE.

Widlar's first design was a bit more complicated, using 14 transistors and producing 5 Volts with four VBEs in series and the delta-VBE multiplied by a factor of about 40. It is no longer used.

There is no such thing as an absolutely precise bandgap voltage. You will find that the voltage at which Vref has no temperature coefficient can be anywhere between about 1.18V and 1.25V due to several effects. First, the bandgap voltage is slightly dependent on the doping level. Second, the bandgap potential of a semiconductor changes with pressure (or stress). And third, we are using (presumably) diffused resistors which have a temperature coefficient of their own. Fourth, as pointed out before, VBE vs. temperature is not an exact straight line; thus Vref vs. temperature will always show a slight upward bow.

Nevertheless, such a bandgap reference voltage can have an accuracy of better than ± 3%, without trimming any of the components.

Apart from base currents (which can be compensated in more advanced designs) there are two main sources of error in a bandgap reference:

1) The VBE. This is an absolute, not a ratio. You have to rely on the precision with which dopants can be introduced into silicon in the process. In a well-controlled process this amounts to about ± 10mV uncertainty, or about 0.8%. Be aware that prototypes from a single wafer (or even a single run) will not give you any indication how much this varies in production over many wafers.

2) Ratios. In Widlar's first bandgap reference there are two ratios of significance: Q1/Q2 and R3/R2 (and also R1/R3). To minimize these errors you simply make these devices large.

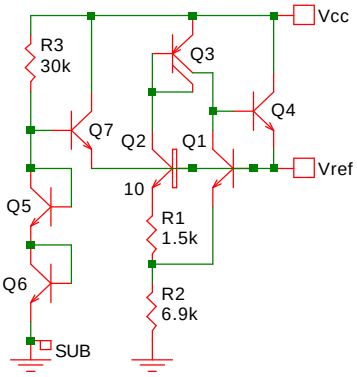


Fig. 7-4: The Brokaw cell.

Four years after Widlar, Paul Brokaw published a paper entitled "A simple Three-Terminal IC Bandgap Reference". The core of the "Brokaw Cell" is formed by Q1, Q2, Q3, Q4, R1 and R2. (His actual circuit contained 14 transistors, so it wasn't so simple after all).

The Brokaw cell needs a start-up circuit, which has been added here (Q7 lifts Vref to one VBE, which is sufficient for Q1 and Q2 to start drawing current).

Q2 has 10 times as many emitters as Q1 (an arbitrary choice, more would be better), so there is a delta-VBE of 60mV (at room temperature) across R1. Q3 is a current mirror, forcing Q1 and Q2 to run at the same current. Q4 completes the feedback loop from the collector of Q1 back to the input bias of the differential pair Q1/Q2 and supplies a moderate amount of output current.

When the circuit is in balance a multiplied delta-VBE shows up across R2. Thus Vref is that voltage plus the VBE of Q1. The value of R2 is selected to achieve a zero temperature coefficient for Vref.

The gain in the feedback loop is limited, which eliminates the need for extra frequency compensation capacitors but results in a relatively high output impedance (about 80 Ohms). Because of the emitter-follower output transistor (Q4), the minimum supply voltage is 2.2V (0 to 100^{o}C), or about 1 Volts above Vref.

Figure 7-5 shows a modification of the Brokaw cell for operation at low supply voltage. Output current is now supplied by Q6, a somewhat larger than normal lateral PNP transistor, capable of delivering 5mA. Q4 forms an additional gain stage, lowering the output impedance to 9 Ohms. Note that the operating current for Q4 is carefully set by Q5 and R3, with R3 having the same value as R2. In this way the base currents of Q3 and Q4 cancel (and, in addition, Q1 and Q2 have identical collector voltages). The minimum supply voltage is now 1.6 Volts.

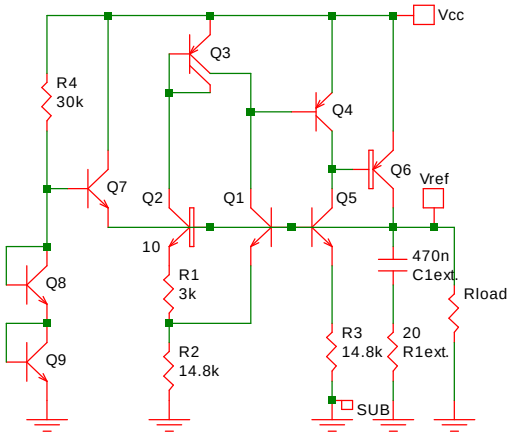


Fig. 7-5: Three terminal Brokaw reference with start-up.

The design procedure for such a bandgap reference is very simple. First you set the emitter ratio of Q2/Q1. The two devices should have identical emitters for best matching, Q2 just has more of them. Make the ratio as high as you can; with a ratio of 2:1 you get a delta-VBE of only about 18mV, which puts a strain on the matching. At 10:1 the delta-VBE is about 60mV (again: at room temperature) and the matching requirements is less severe. At 50:1 the delta-VBE amounts to about 100mV at which point matching becomes easy (you also have a large number of emitters which, statistically improves matching).

With the emitter ratio chosen, you now know the value of delta-VBE appearing across R1. You then set R2 so it drops about 600mV; in this particular case twice the current flows through R2 as flows through R1, so a 5:1 resistor ratio will give you 10:1 voltage ratio.

Next comes the simulation, and for this you need good models, including the temperature coefficient of the resistors. Plotting Vref against temperature, you will almost certainly see a marked temperature coefficient.

Now simply change the value of R2 until this temperature coefficient is zero, end to end. A higher value for R2 will give you a more positive tempco.

Ideally, R1 and R2 should have a ratio so that you can divide them into identical sections in the layout. In this example 3k/15k would be perfect, breaking the divider into six identical sections of 3kOhms each. In reality this rarely happens. You may find that, by changing the value of R1 (thus drawing more or less current) you can get to this ideal ratio, but if you don't there is a compromise: Use a smaller basic section (say 750 Ohms) and then get the odd value of R2 by making the last section (or perhaps the last few sections) a combination of parallel and series connections of the basic resistor element.

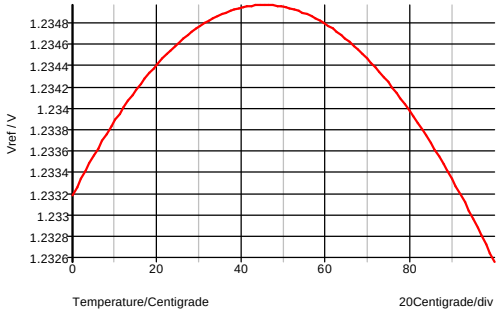


Fig. 7-6: Characteristic bow in the temperature curve of a bandgap reference.

Vref shows the characteristic bow of a bandgap reference, due to the slight curvature of VBE. This amounts to about 0.18%.

This curve was obtained using models for a simple 5-Volt bipolar process. The results are going to be different for other processes, you will need to find the optimum value for R2 using your own models. Also, the final value for Vref will most likely be somewhat different too.

When you plot Vref vs. temperature from a simulation you get a false sense of precision. You will see the curve of figure 7-6 only once in a while on a real IC, one that happens to have the exact nominal parameters. What you have to live with is more like figure 7-7, obtained from a Monte Carlo run. Over a range from 0 to 100^{o}C the variation is about ± 2.5%. This can be reduced by trimming and the best component to trim is R2. As you can see there is a distinct relationship between

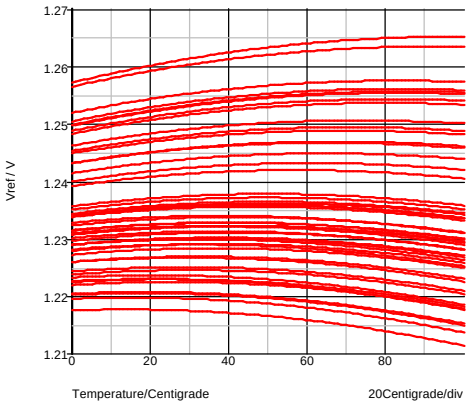


Fig. 7-7: In production (Monte Carlo analysis) you will see a larger deviation than merely the curvature.

voltage and temperature coefficient. R2 controls both.

Even with trimming, there is a limit to accuracy. When a chip is attached to a lead-frame in a package, there is always some stress. Stress changes the bandgap potential and, unless some unusual precautions are taken, Vref can change as much as 0.5% (in either direction) compared to the value measured (or trimmed to) on the wafer. This can of course be avoided if the reference can be trimmed in the package.

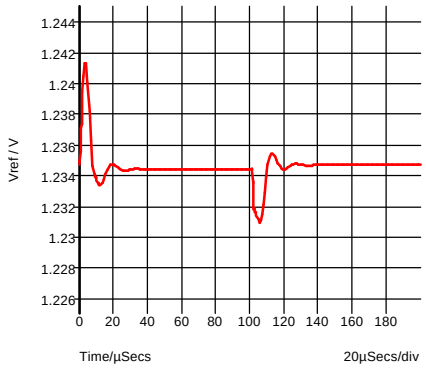


Fig. 7-8: Pulse response, indicating stability.

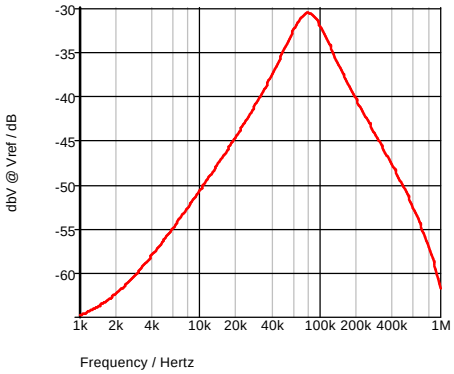


Fig. 7-9: Power supply rejection.

The use of a PNP transistor at the output makes frequency compensation of a feedback loop difficult. This is true especially for a slow lateral one. About the only practical way to compensate this reference is to place a large (i.e. external) capacitor at the output. Even so a small resistor in series with the capacitor is required to create a zero (see chapter 6). The loop is stable but the power supply rejection at 100kHz is a mere -30dB.

It's Widlar's turn again. Four years after Brokaw he came up with a whole series of new bandgap reference designs. Figure 7-10 shows one of them. Q1 and Q2 have a 4:1 emitter ratio (just to show some variety) and their emitters are connected together. So the delta-VBE shows up between their bases, i.e. across R2. This amounts to:

deltaVBE = ln4*26mV = 36mV

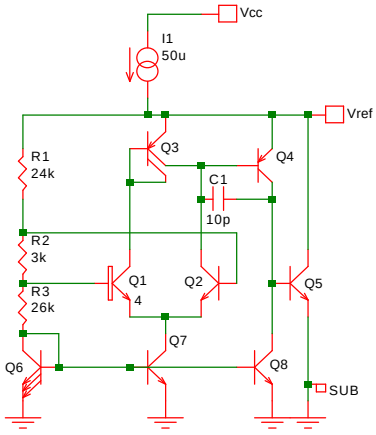


Fig. 7-10: Another Widlar bandgap reference.

at room temperature. Since there is only one current flowing through all three resistors (save the base currents of Q1 and Q2) the voltage drop across all of them is 36mV(53k/3k) or 636mV. Add the 600mV VBE of the diode-connected Q6 to this and you get a temperature compensated reference voltage of 1.236V. (Again, this value and the required values for R1 or R3 may be somewhat different for other processes).

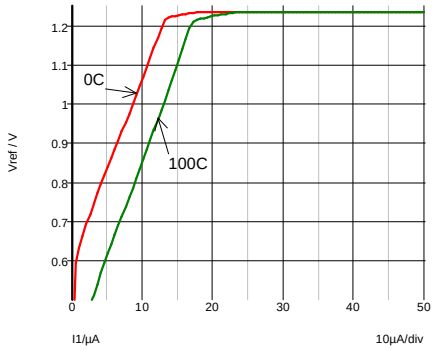


Fig. 7-11: Minimum operating current.

The multiplying resistor as been split into two parts (R1 and R3) to provide enough headroom for the transistors to operate. The operating current for the differential stage (Q1, Q2) and the second stage (Q4) is reduced by making Q6 three times as large as Q7 and Q8.

This reference requires a minimum current of 25uA to operate properly. Above that level the impedance at the output is about 10 Ohms. Frequency compensation is easily accomplished by enhancing the Miller capacitance of the slowest device, Q4, with a 10pF capacitor.

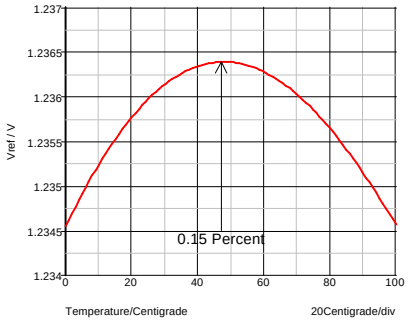


Fig. 7-12: Deviation over temperature (the bow.)

Let's look at the variation again.

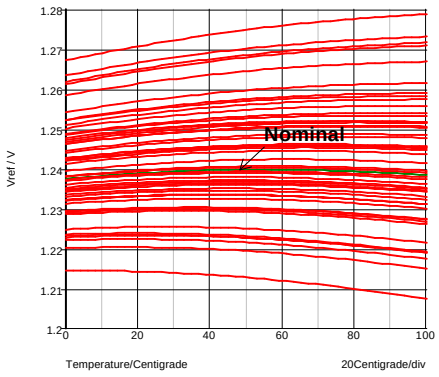


Fig. 7-13: Again the bow is a minor factor in the overall variation.

Once R1 (or R3) is optimized for near-zero change at the temperature extremes, we get the inevitable bow. For this reference it amounts to 0.15%.

When we put this bow in context, namely add to it the production variations due to the absolute value of the VBE and the matching variations of the resistors and transistors, we get quite a different picture. The 0.15% bow is overwhelmed by the ± 3% overall variation. (The variation, however, can be reduced to perhaps ± 2.3% by choosing a larger emitter ratio).

At the same time Widlar introduced his new designs he also came up with a way to reduce the bow, a method which is now called **second-order temperature compensation**.

To illustrate it we use the same bandgap reference again, with one transistor added. A portion of the voltage across the resistor string is tapped by the base-emitter diode of Q9 with a large-value resistor. The tapped voltage has a positive

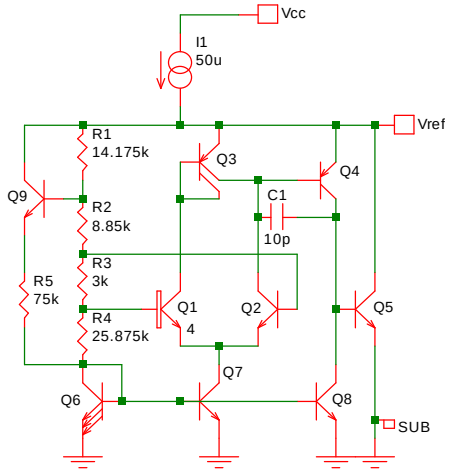


Fig. 7-14: Bandgap reference with curvature correction.

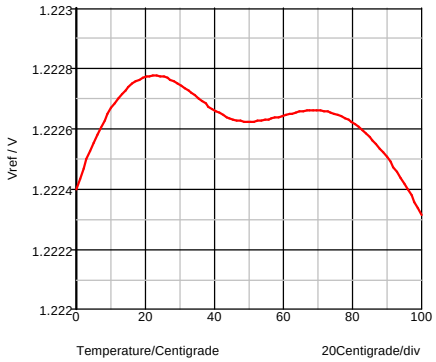


Fig. 7-15: The bow reduced by second-order temperature compensation.

temperature coefficient, the base-emitter diode a negative one. At about 40°C Q9 and R5 start feeding a small current into Q6, which increases as the temperature is increased. This bends the right-hand side of the characteristic bow upward. R1, R2 and R4 then need to be adjusted to level the curve, a somewhat delicate process. The net result (after several adjustment cycles) is a flatter curve, showing a deviation of just 0.04%.

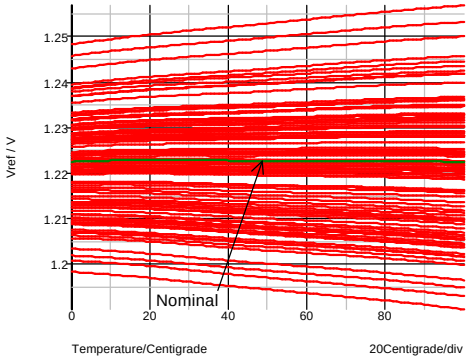


Fig. 7-16: The overall variation now overwhelms the remaining bow, so this approach should only be used with trimming.

But let's put this in context again. We may have straightened the nominal curve, but it is still subject to the variations caused by VBE and matching. Adding this (Figure 7-16) we see little or no improvement in overall accuracy. For this reason second-order curvature correction only makes sense if a bandgap reference is trimmed in

a very sophisticated way, reducing production variation to considerably less than 1%.

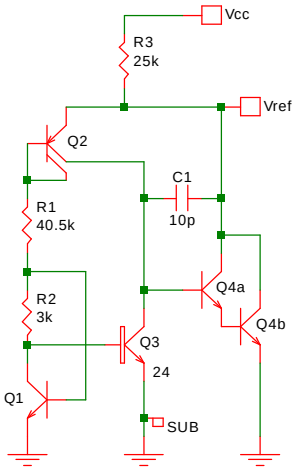


Fig. 7-17: A different design for a bandgap reference (2.5V).

Here is a more modern bandgap reference which is more accurate (without trimming) than the previous examples. Just 4 transistors are used, one with a dual base/emitter. It is basically a two-terminal reference, fed by R3. There are two VBEs in series, so the output voltage is twice the bandgap potential, 2.45V. With R3 = 25kOhm the optimum Vcc range is 4.5 to 5.5V.

The delta-VBE appears across R2, given by the 24:1 emitter ratio of Q3 to Q1 (about 83mV at room temperature) and is multiplied by R1 to about 1.2 Volts. The difference here is the placement of R2 in the collector circuit of Q1, thus subtracting rather than adding the delta-VBE. One VBE is that of the lateral (split-collector) PNP transistor Q2, the other the NPN transistor Q1. Lateral PNP transistors generally have a narrower variation in VBE, but work only over a limited current range.

The error signal is picked up by a Darlington transistor (Q4, one collector region, two base-emitter patterns).

Variation in production over a temperature range of 0 to 100^{o}C is a mere ± 1.6%. As always, the values given here are for a specific process, with fairly large dimensions (the resistors are 4um wide). You may need to adjust R1 for other processes (and certainly for other emitter ratios).

The circuit is stable with a load capacitance of less than 50pF or greater than 200nF. With a 330nF capacitor at Vref power supply rejection is -60dB, increasing further above 10kHz. The output impedance is 25 Ohms. The circuit is intended as a reference only, but it can sink several milliamperes. If more sourcing current is needed, you simply decrease the value of R3.

It is possible to modify this circuit for 1.2 Volts but, as a consequence the performance suffers a bit (which is almost always true when you move to lower voltages).

In figure 7-18 only a single diode-connected transistor (Q1) is used. A second one mirrors one-third of the current, which is compared with the mirrored current of Q4. Here the emitter ratio is 20:3. A second stage (Q5) increases the loop gain, lowering the output impedance to about 1.7 Ohms. R3 is optimized for operation from 3 to 3.6 Volts, consuming 90uA.

Production variation from 0 to 100°C is ± 2.2%.

Frequency compensation is a bit tricky. With a load capacitance of 500pF or less the circuit is stable and has a power supply rejection of -80dB below 10kHz, -60db at 100kHz and a peak of -40dB at 1MHz.

Figure 7-19 shows the same circuit, transformed into a 3-terminal reference, or a mini voltage regulator. It uses an NPN transistor to supply the output current, which delivers a greater current than a (lateral) PNP transistor and makes frequency compensation an easy job, but only works down to 2.2 Volts supply voltage.

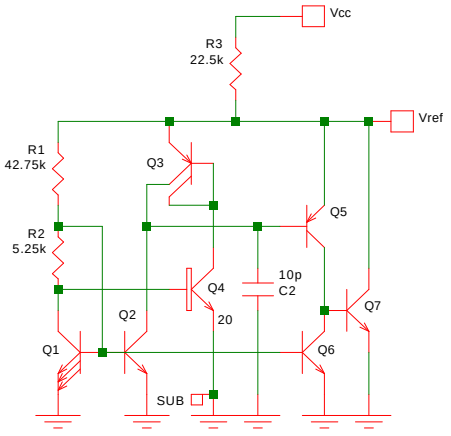


Fig. 7-18: A similar circuit, but with a single VBE (1.25V).

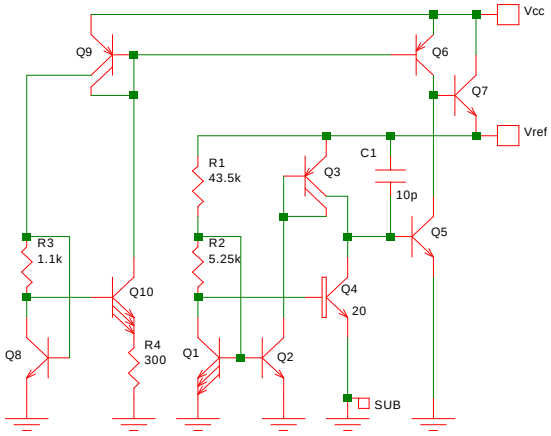


Fig. 7-19: Three-terminal version of figure 7-18.

The base current for the output transistor is supplied by an independent current source consisting of Q6, Q8, Q9 and Q10. This is the self-starting current source discussed in chapter 5, figure 5-12. The last transistor of the bandgap reference, Q5, diverts the unneeded current from Q6. Q6 supplies about 100uA. With a maximum hFE of Q7 (at high current) of 100, the output current is limited to about 10mA (depending on the size of Q7), which prevent burn-out.

Production variation (3-sigma) over a temperature range from 0 to 100°C is ± 2.4%. The output impedance is 1.5 Ohms and the circuit is stable with any load capacitance.

Low-Voltage Bandgap References

The principle followed in the bandgap references so far has been this: add two circuit elements with equal but opposite temperature coefficient, so that the sum of the two has a temperature coefficient of zero.

One of the circuit elements is a diode, which has a voltage drop of about 600mV. The second circuit element is a multiplied delta-VBE, which also amounts to about 600mV. Therefore the minimum reference voltage possible is about 1.2 Volts.

This is only true if we *add* these two voltages. There are other approaches which avoid addition. Let's look at two of them.

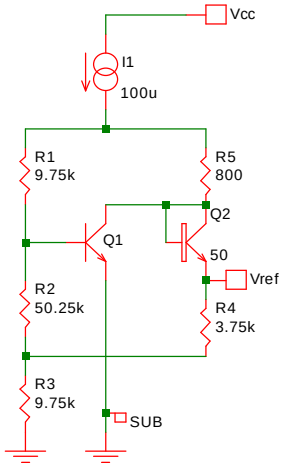


Fig. 7-20: 200mV Vref design by Widlar.

Here is the ferment mind of Bob Widlar again. As early as 1978 he suggested a circuit which works down to 1 Volt supply, a single battery.

First off he uses just about the largest emitter ratio practical, 50:1 between Q2 and Q1. This gives a delta-VBE of about 100mV at room temperature.

The VBE appears at the base of Q1 to ground. Thus the voltage at the entrance of the current source is higher by a fraction of a VBE. Now assume R5 to be zero. Thus the voltage at Vref is that fraction of a VBE plus the delta-VBE of the 50:1 ratio in emitters of Q2 and Q1. If R1 is dimensioned such that the fraction of the VBE amounts to about 100mV, then we have a temperature-stable Vref of 200mV.

R5 provides some compensation for changes in I1 and connecting R4 to a tap at R2/R3 rather than ground creates a minor amount of second order temperature compensation.

A Vref of 200mV is just about the maximum value you can get from this design. Even with a much larger emitter ratio, say 200:1, delta-VBE only amounts to 138mV, i.e. Vref would be about 272mV.

The second approach is considerably more complex but has greater flexibility. Two *currents* are created, one with a positive temperature coefficient, the other with a negative one. Summed, they produce a voltage drop in a resistor and this voltage drop has a near-zero temperature coefficient.

The first current depends on the 3:1 emitter ratio of Q6 and Q4 and the fact that Q4 runs at twice the current compared to Q6. Thus the effective emitter ratio is 6:1 and the current is determined by the delta-VBE (47mV at room temperature) and R1. The feedback loop has a gain of 3, carefully controlled by the 3:1 emitter ratios of Q1/Q3, in this way the loop is frequency-compensated by the device capacitances. The loop is self-

starting by leakage currents and the collector currents of Q2 and Q5 have a positive temperature coefficient. Two identical currents are derived by Q7, one feeding the output resistor R3, the other starting the second current source.

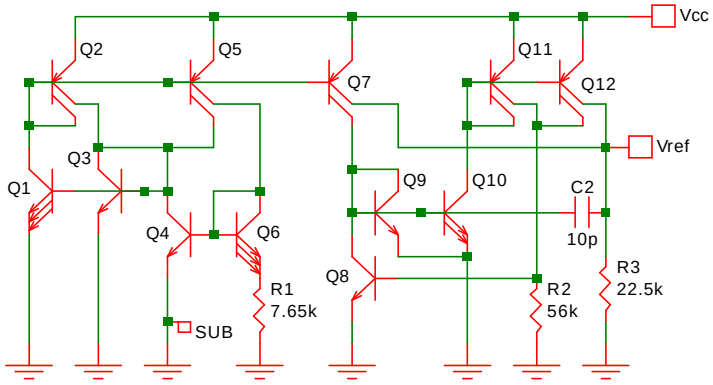


Fig. 7-21: Bandgap reference with a minimum supply voltage of 0.9 Volts.

The second current depends on the VBE of Q8 and the value of R2. Again, the loop has a limited and well-controlled gain (the emitter ratio of Q10/Q9 and the 2:1 collector ratio at Q11/Q12), but a small frequency compensation capacitor is still required. The collector currents of Q11 and Q12 have a negative temperature coefficient and one collector of Q12 feeds the output resistor R3. The sum of the two currents flowing through R3 cause a voltage drop of 250mV, with a temperature coefficient near zero.

The two currents can be adjusted independently with the values of R1 and R2, allowing fine-tuning of the temperature coefficient. The magnitude of the output voltage can be selected with the value of R3 without affecting the temperature coefficient.

Note that the currents depend on the resistor values. They will vary in production but R3 tracks these variations and the output voltage depends only on resistor matching.

The output impedance is that of R3. Unless the load draws only a very small current you will need an output buffer.

This bandgap reference works down to 0.9 Volts supply and the change in output voltage from 1 to 1.5V Vcc is 0.25%. Power supply rejection is -55dB up to 10kHz. To keep this low at higher frequencies you will need an external capacitor (10nF) at the output.

Production variation is ± 3.6% from 0 to 100^{o}C, which illustrates that the lower the supply voltage the more difficult it is to get high performance, even if a more elaborate circuit is used.

CMOS Bandgap References

Let's face it: a bandgap reference is a bipolar concept. It needs a diode and the difference between two diodes. And the only diodes good enough are diode-connected bipolar transistors (or, in some designs, the base-emitter diodes of bipolar transistors).

Fortunately there are some layers in a CMOS integrated circuit which, although not intended for this purpose, can be used to make a passable bipolar transistor. The most obvious ones are those used for a p-channel transistor, the p-type region (source, drain) forming the emitter, the surrounding n-well the base and the substrate the collector.

Such a device has limitations. First, the collector is permanently tied to the lowest supply voltage. No flexibility there at all. Second the gain (hFE) is very low, about 7. In a bipolar process we rely on the high gain (at least 100) to effectively eliminate the base resistance as a source of error. So the CMOS substrate PNP transistor only works if we make it large (which we probably want to do anyway to get reasonable accuracy).

It is also possible to make lateral PNP transistors in CMOS, using the p-channel drain/source diffusions as both the emitter and the collector. Such devices have a reasonable gain (100 or so) but, unlike the substrate devices, they are hardly ever characterized by the foundry, which means you can't consider them unless you want to spring for a rather expensive evaluation run.

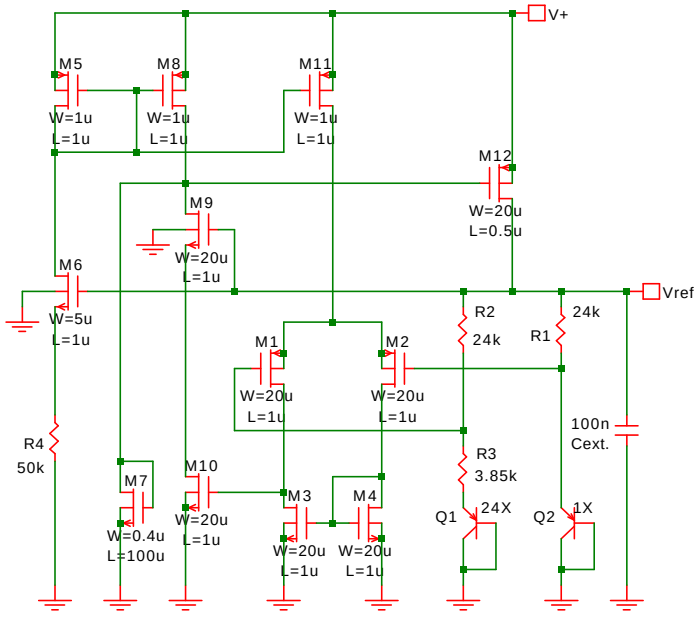


Fig. 7-22: CMOS bandgap reference with substrate diodes.

For this reason we will consider only a CMOS bandgap reference using substrate PNP transistors here. Q2 has a single (10um x 10um) emitter, Q1 has 24 of them. Q2 is usually in the center, surrounded by 2 rows and columns of identical Q1 devices. Get used to it: this pattern is very large

compared to a 0.12u, 0.18u or even 0.35u CMOS device.

The delta-VBE appears across R3, with R1 and R2 being equal. The error voltage is amplified by M1, M2, M3, M4, M10 and M12. These devices need to be as large as indicated. For M1, M2, M3 and M4 the prime requirement is matching, which gradually improves as the area (channel length times width) is increased. (Keep in mind that we are working down at a level of a delta-VBE, which here amounts to about 82mV). For M10 and M12 the width needs to be substantial to get sufficient gain (transconductance) and an increased length helps to reduce the influence of power supply variations. To reduce this even more M9 (a cascode stage) has been added.

M7, a narrow and very long transistor starts the circuit by feeding a small current into the loop. Once sufficient voltage appears at Vref, M6 and R4 take over and supply the operating current, mirrored by M5, M6 and M11.

M12 is a p-channel transistor, which provides a low minimum supply voltage (1.5V) but, as we have seen before, make frequency compensation difficult. The only practical way to do this is with an external capacitor, though placing it at the output also provides for a good power supply rejection (-60dB). The output impedance is 0.5Ohms up to about 1mA.

With the transistor sizes as shown and the resistors 4um wide you can expect a production variation of ± 1.8% over a temperature range from 0 to 100^{o}C.

A word of caution: A bandgap reference is the ultimate test of accuracy for device models. For example, it is very difficult to measure a VBE over temperature accurately enough on a wafer so that it will predict the behavior of a bandgap reference. With most processes you need to make a bandgap reference to verify the models.

8 Operational Amplifiers

Op-amp design is a specialty, peopled by a small group of engineers forever dedicated to the quest of finding the universal building block. None has ever been found (hence the large number of different op-amps) but still they toil. Year after year they come up with small improvements; and each new design has one imperative requirement; it must work in any application without creating smoke or - heaven forbid - oscillation.

When designing an op-amp for an ASIC the precise application is known, so the circuit does not need to be universal and the task is easier. Not exactly a cinch, but nothing compared to what a designer of commercial op-amps has to face.

The majority of op-amps have three stages. The first stage converts the differential signal into a single-ended one; the second one provides the bulk on the gain and the third one the required output power. There is no law that says it has to be this way, it just turned out to be an approach that works well.

Bipolar Op-Amps

In our first circuit Q1 and Q2 form the differential pair and Q3, a split-collector lateral PNP transistor, is connected as a current mirror or active load. At the collectors of Q2 and Q3 we have a high impedance, limited only by the base current of Q4 and the Early effects in Q2 and Q3. The second stage, Q4, has as a load the current sink Q8 and the base current of Q5, thus its gain is also limited only by those two. The output stage is a simple emitter follower with a pull-down current sink.

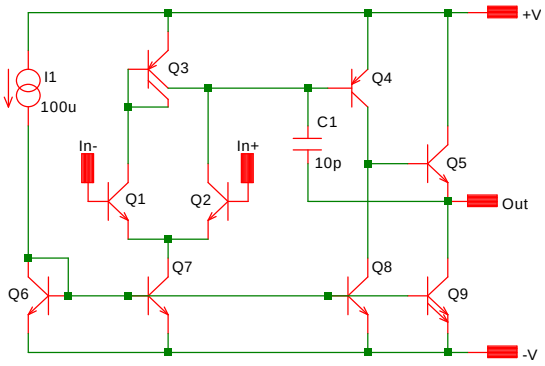


Fig. 8-1: Simple 3-stage op-amp.

The operating currents for all three stages are derived from I1 through a multiple current mirror. Q9, with two emitters, delivers two times I1, an arbitrary choice; more emitters could be used if more current is required at the output for negative-going signals.

Q7 and Q8 are identical, which is no arbitrary choice. By making the collector currents of Q4 and Q3 (both halves) equal, Q3 takes as much base current from Q1 as Q4 does from Q2. Thus there is no designed-in base current error and the offset at the input is zero (for ideal matching).

Common Mode Range: Describes the minimum to maximum DC level at which the two inputs are functional. Desirable is **rail-to-rail**, i.e. from the negative supply to the positive one, but many op-amps only work with the inputs a volt or two above -V (or ground for a single supply) to some voltage below +V.
Common-Mode Rejection: If you connect both inputs together, bias them at a functional DC level and superimpose on the bias a small AC voltage, no signal should ideally appear at the output. In reality a small signal leaks through and the measure describes how much smaller this signal is (in dB) compared to the input.

The common-mode range in this design has limitations. The two inputs need a dc level of at least one VBE (Q1, Q2) plus a saturation voltage (Q7) above -V. If they move below this level the input pair gets no operating current. Also, the inputs need to be about 200mV below +V, otherwise Q1 or Q2 saturate and Q3 or Q4 are without current. At the output Q5 can pull the output only to within a VBE of +V.

There is also another flaw: being bases of bipolar transistors, the inputs need a current. With a minimum hFE of 100 and 50uA flowing per transistor, this amounts to base current of 0.5uA worst case. With a 100kOhm input resistance for one input and zero for the other this could amount to as much as a 50mV error at the output.

Frequency compensation is achieved with a single capacitor from the output to the high-impedance node at the output of the first stage. It could have been placed just from collector to the base of Q4, but the simulation shows a small advantage for the shown configuration.

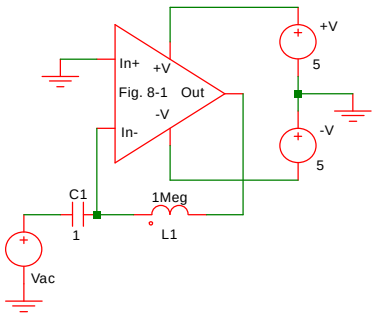


Fig. 8-2: Simulating loop gain and phase.

Let's first double-check this frequency compensation. To do this (as explained in more detail in chapter 6) we place a very large inductor in the feedback loop and feed an AC signal into the input through a very large capacitor. The inductor (1MH, i.e. 1 million Henry) provides the DC bias to the input but blocks

the AC signal. You want to choose the inductor and capacitor large enough so they have no effect at even the lowest frequency of interest for the circuit. Neither 1 million Henrys nor 1 Farad are practical values; they don't need to be.

And here is what we get: The open-loop gain is about 85dB. The dominant pole, given by the compensation capacitor, is at about 2kHz, after which the gain decreases steadily and reaches unity (0dB) at about 12MHz. At that point the phase is still at about 50 degrees, marginal but probably adequate.

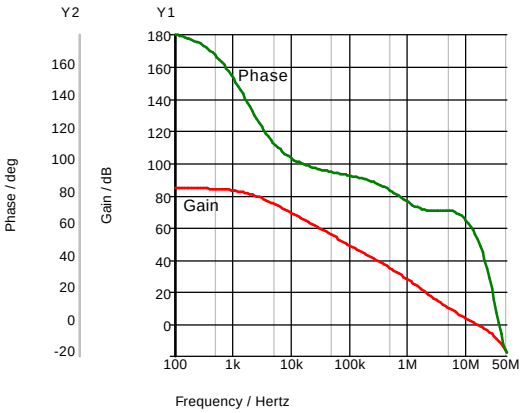


Fig. 8-3: For stable operation the gain of a loop must reach 0dB before the phase reaches 0 degrees.

But, as pointed out in chapter 6, a phase-margin analysis is not the real test of stability. AC analysis uses infinitely small signals (even if it says the signal is 1 Volt) and the operating currents and voltages are not disturbed. So, to be certain, we would have to repeat this analysis for the DC conditions over the entire (large-signal) range of the circuit, a tedious task at best. We get a more immediate picture by observing a large-signal pulse.

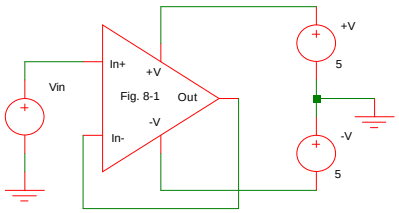


Fig. 8-4: Buffer connection.

To do this we eliminate the inductor and capacitor in the feedback path and connect the signal to the input. Without resistors in the feedback path, this is a buffer connection, i.e. a closed loop gain of one; with the entire open-loop gain (85dB) being judged, this is the most severe test for stability.

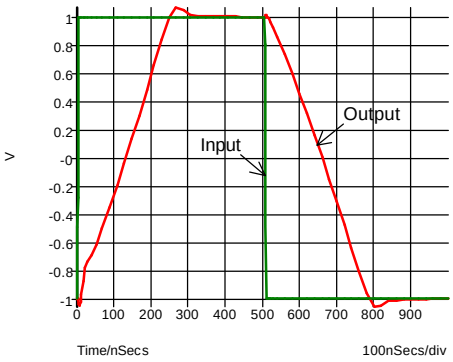


Fig. 8-5: Input and output waveforms for the buffer connection.

The output waveform shows the amplifier to be very

stable, there is no ringing, just a small overshoot.

But the curve shows something else: rise and fall-times are substantial and almost straight lines. This is the **slew-rate**, the time it takes to charge and discharge the 10pF capacitor over a 2-Volt span with the operating current. You can speed it up by increasing the operating current, at the cost of power consumption.

In doing this test we assume that the op-amp needs to be operated as a buffer. What if, in a specific application, the closed loop gain is never lower than 40dB? In such a case we only have an excess open-loop gain of about 45dB, which makes compensation considerably easier.

Let's examine this. In figure 8-6 we have the same circuit as in the loop gain analysis before, except that the feedback resistors are in the loop also.

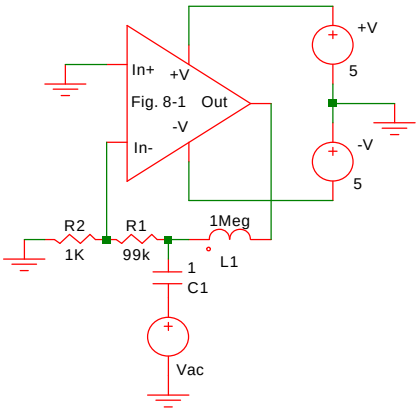


Fig. 8-6: Simulation of loop gain and phase with closed-loop gain of 100.

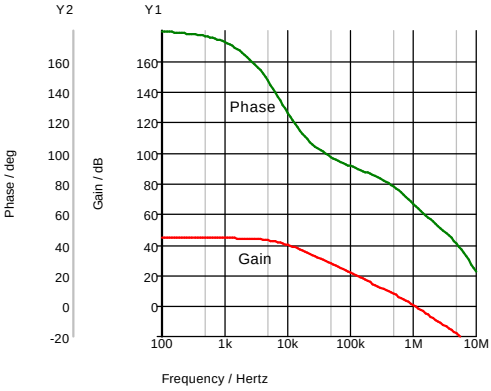


Fig. 8-7: With the lower gain the loop shows greater phase margin, even though the compensation capacitor (C1 in figure 8-1) is reduced to 2pF.

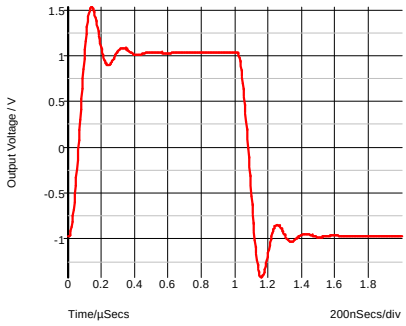


Fig. 8-8: Pulse response with 40dB loop gain and a 2pF compensation capacitor.

The gain now shows a 45dB maximum instead of 85dB and thus drops to 0dB at a lower frequency. We can now reduce the value of the compensation capacitor from 10pF to 2pF and still have a phase margin of over 60 degrees.

When we check the stability with a pulse (with only the resistors in the feedback loop and the signal connected to the positive input) we see that the amplifier is just stable enough (fewer than 4 peaks in the damped

oscillation). And, because the compensation capacitor is smaller, the slew-rate is substantially higher.

No consideration has been given so far to noise and noise performance is in fact not that great in this design. The majority of the noise is created in the input stage; in subsequent stages the signal is larger and the influence of noise is correspondingly reduced. To lower the noise in Q1 and Q2 the devices need to be larger and their operating current higher.

Figure 8-9 shows the noise performance of this amplifier in the buffer configuration (since the gain is 1, output noise and input noise are equal). As you can see, lowering the operating current *increases* noise, an unpleasant fact of life.

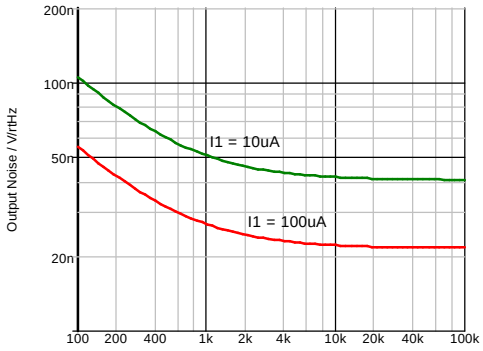


Fig. 8-9: Noise of the amplifier in the buffer connection.

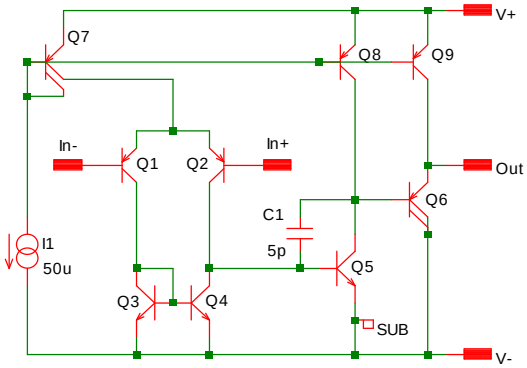


Fig. 8-10: PNP-input equivalent of figure 8-1.

Naturally you can invert the polarity of the transistors and design an op-amp whose input can operate close to the negative supply (but loses its ability within about 1 Volt of the positive supply). Figure 8-10 is the PNP-input equivalent of figure 8-1, still with a rather primitive output stage (Q6, Q9), which can pull the output to within about 150mV of the positive supply (if the load does not require more than 50uA), but only down to about 1 Volt above the negative rail.

Let's now consider a design in which the inputs can be operated all the way down to the level of the negative rail and the output swings (almost) rail-to-rail. The input stage in figure 8-11 is a configuration known as the **folded cascode stage**. With an operating current (I1) of 10uA the voltage drop across R1 and R2 is a mere 50mV, thus the inputs can go about 250mV *below* the negative supply rail without saturating Q1 or Q2.

The collector currents of Q1 and Q2 upset the balance of the Wilson current mirror Q3, Q4 and Q5 and the difference signal is picked up by Q6.

The output stage has two branches to it. The first one is simply Q14, a grounded emitter amplifier. All other transistors in this block serve its antipode, Q13.

Note the three diode connected transistors Q7, Q8 and Q9. They set a voltage for the base of Q11. If you follow the emitter-base junctions of Q11, Q12 and Q14, you notice there are also three diodes in series to the V- rail. Thus, as the input signal to the output stage moves up and down (by a few millivolts), the current in Q11 fluctuates. It is this current, amplified by the size ratio of Q13 to Q10 (here about 6) that becomes the pull-up portion at the output. Q7 to Q9 are deliberately made larger than Q11, Q12 and Q14 so that the idle current in the output is small. This creates a small "dead-band" (see chapter 16) but, because of the large loop gain the distortion is very small (0.0004% for a ± 4.7Vp signal at 1kHz).

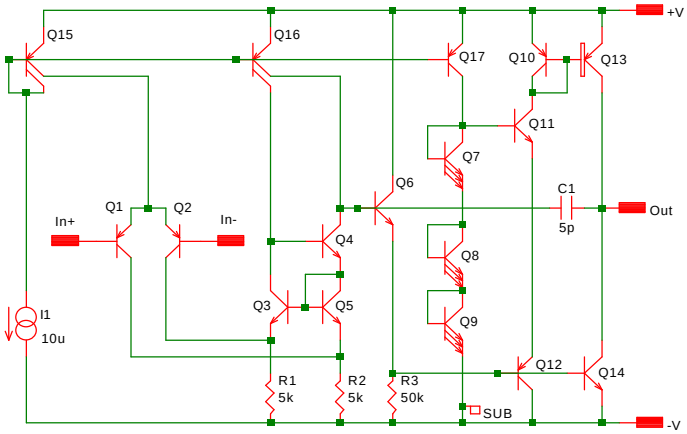


Fig. 8-11: Op-amp with folded-cascode input stage and (almost) rail-to-rail output.

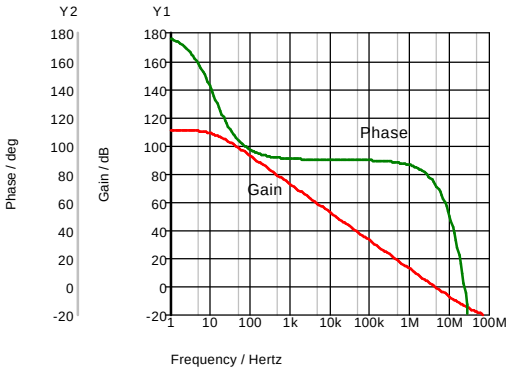


Fig. 8-12: Phase margin of figure 8-11.

In this circuit we are fortunate to find a node ideally suited for the connection of a compensation capacitor to the output: at the base of Q6 the signal has a phase opposite to that at the output; the base of Q6 and the collectors of Q4 and Q16 all represent a high impedance; and there is substantial voltage gain between it and the output. Which all says that this op-amp can be compensated (at unity gain) with a single 5pF capacitor, even though the loop gain is 110dB. Not all designs behave that well.

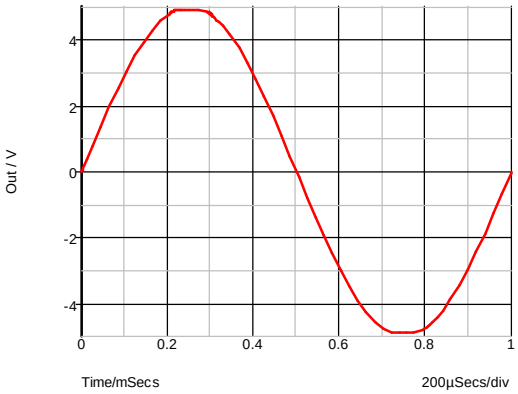


Figure 8-13: Output swing is limited by the saturation voltages of Q13 and Q14.

Though the inputs can work at the level of V- (or ground if you have only a single supply), the output cannot. It is the sad truth for a bipolar transistor that there is a saturation voltage. Unlike an MOS transistor, which is simply a voltage-controlled resistor, the bipolar transistor is the interaction of two junctions with different doping levels and sizes. Even when fully turned on there is a minimum voltage drop of about 150mV between emitter and collector in transistors Q13 and Q14. Thus the output can never be at the rails, only approach them.

The use of lateral PNP transistors at the input and the low operating current is not kind to noise: 27nV/rtHz at 10kHz and up (white noise). At 1Hz the flicker noise rises to 80nV/rtHz. If you have vertical PNP transistors at your disposal and can afford a higher operating current, these figures drop by a large factor (but you need to carefully re-simulate the entire circuit; frequency behavior is bound to be entirely different).

Another unsatisfactory parameter is the input current. Each input transistor runs at 5uA; with a minimum hFE of 100 (in a good process) the base current can be as high as 50nA. But there is a solution to this: more transistors.

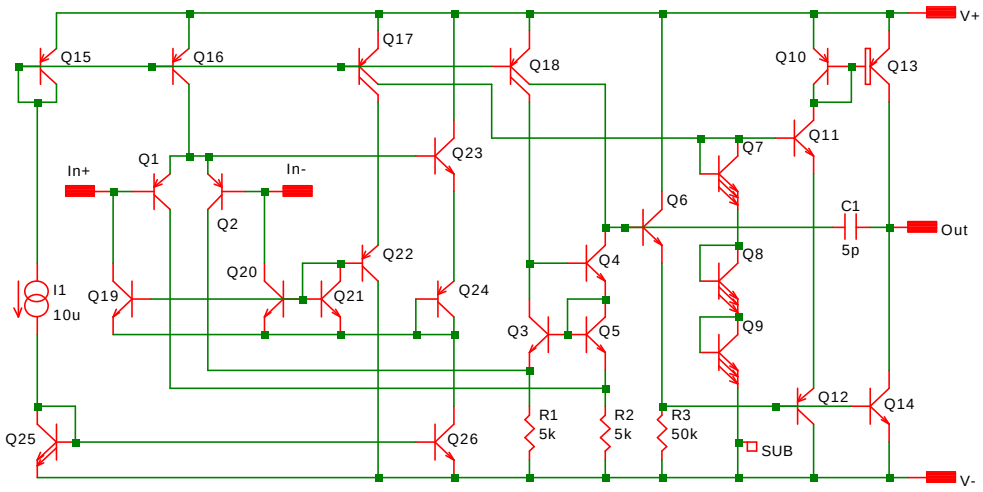


Fig. 8-14: Op-amp of figure 8-11 with base-current compensation for the input stage.

In figure 8-14 eight transistors are added. Their job is to pull as much current out of the bases of Q1 and Q2 as they naturally require, so that he external circuit does not have to do this.

The key is Q22. It is identical in size and design to Q1 and Q2, has the same operating current and very close to the same collector-base voltage (created by the base-emitter voltage of Q23 and the diode-connected transistor Q24). Thus its base current must be the same as those of the input transistor. This base current is mirrored by Q21, Q20 and Q19 and fed to the inputs, *opposing* the two base currents.

The cancellation of the base currents is never perfect of course, the current levels are too small to get precision matching. But the net input currents are down to 2nA, a 25:1 improvement.

In the last bipolar op-amp the goal is not an ultra-low input current, but low-noise performance with a reasonably low input current.

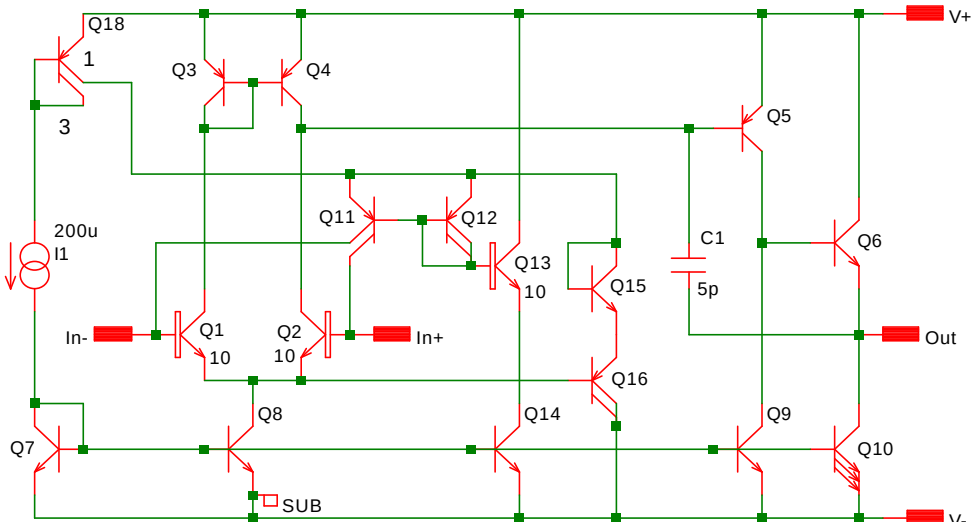


Fig. 8-15: Bipolar op-amp optimized for low noise.

This circuit is almost identical to that of figure 8-1. The input transistors are now again NPN, but with the base-current canceling scheme added. The key transistor is Q13; since the hFE of an NPN transistor changes much less with current than that of a PNP device, we can afford to run it at twice the current and then divide the base current by two in the current mirrors Q12/Q11. This brings the input current down to 20nA.

The operating current is much higher than that of the previous circuit and the input transistors are large, which lowers the white noise to 5nV/rtHz.

CMOS Op-Amps

CMOS devices have two advantages for op-amps over bipolar ones: there is no input current (at least not at DC) and, when the transistor is fully turned on, there is only a simple resistance between drain and source (not some complex cancellation of two junctions, resulting in an offset voltage *and* a resistance).

Let's again first look at a simple design. As in figure 8-11 a "folded cascode" input stage is used, this time using N-channel transistors (M5, M6). The primary current, I1, is mirrored with M1-M4 and then again in M7-M10, using the circuit of figure 3-24, so the operating current for the input pair is a constant 20uA.

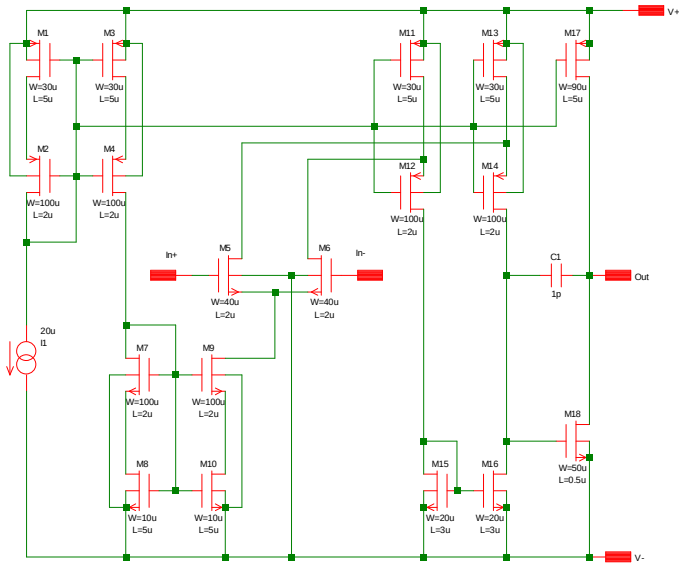


Fig. 8-16: Op-amp with folded-cascode input stage and a simple (and limited) output stage.

The four transistors M1-M4 also steer M11/M12 and M13/14, producing two more accurate currents of 20uA each. The drains of the input transistors are connected to the sources of M12 and M14, which have a potential about 200mV below V+; thus the inputs can operate up to (and about 100mV above) the positive supply.

At balance (In+ = In-) the input pair diverts half of the 20uA current produced in M11 and M13, i.e. M12 and M14 are left with only half the current, about 10uA each. The current out of M12 is mirrored in M15/M16 and opposed to that flowing out of M14. With a large input signal the two currents become unbalanced and each can vary between zero to 20uA. The voltage created by this unbalance is amplified by the output stage (M18 and a simple pull-up current source, M17). The idle current of the output stage is set by the ratio of the channel widths of M1 to M17, i.e. 60uA.

The circuit is compensated with a single 1pF capacitor, utilizing the Miller effect of M18. This works well as long as the load capacitance is small. At 10pF the stability is marginal when connected as a buffer; if the closed-loop gain is never lower than 10 however, the op-amp is very stable, even with a load capacitance as high as 50pF.

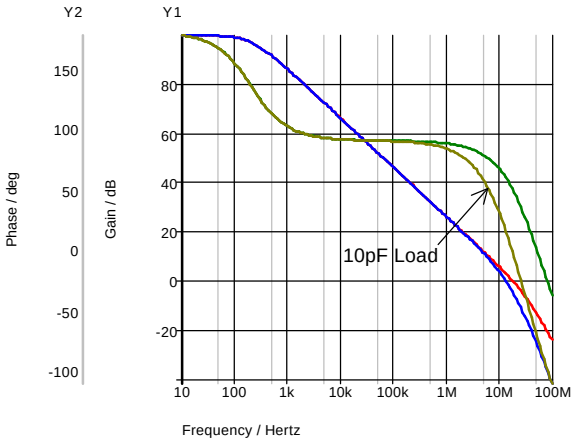


Fig. 8-17: Phase margin becomes critical with a capacitive load, unless the closed-loop gain is 40dB or larger.

With a load resistance of greater than 25kOhms (60uA) the output can move rail-to-rail (or, more precisely, to within about 100mV of each rail).

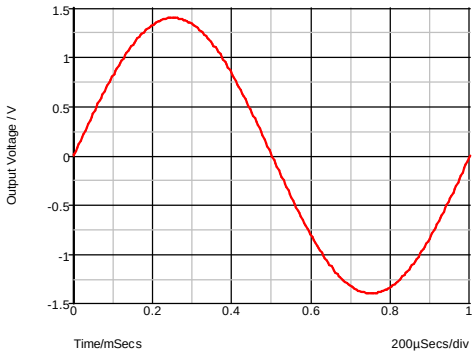


Fig. 8-18: Output swing.

All CMOS examples in this chapter assume a split power supply of 3 Volts total, or ±1.5V; they are operational down to ±0.8V, though with reduced performance.

You need to be aware of the changing open-loop gain. At ground level it amounts to 100dB. As the output moves close to either supply there is a marked drop (66dB at -1.4V, 60dB at +1.4V). This can be improved by making the output devices larger.

What cannot be improved is a fundamental dependence of loop gain on the load impedance. The lower the load, the lower the loop gain.

The input stage has a limited common-mode range. It will work up to about 100mV *above* the positive rail, but not below about -0.8V, i.e. about 0.7V *above* the negative rail.

Let's convert the wimpy output stage into a true rail-to-rail one, with some current capability:

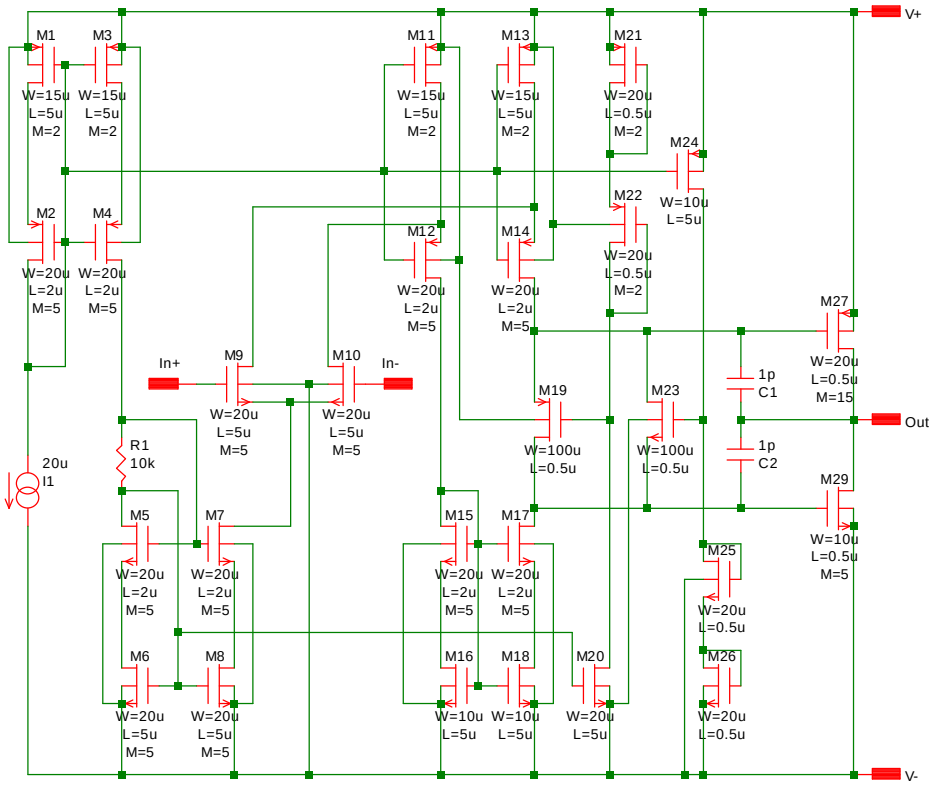


Fig. 8-18: Op-amp with a more capable rail-to-rail output stage.

The trick in designing a rail-to-rail output is in the biasing of the two output transistors. You want a small but well-controlled idle current to minimize any uneven behavior as the output signal is switched from one transistor to the other. In this circuit there are eight transistors whose only job is to set this idle current.

Follow M25 and M26, two "diode-connected" n-channel devices, fed by the current source M24. At the gate of M23 we then have a DC potential of about 1.2V above the negative rail. There is a second path from this node to V-, through M23 and M29, also n-channel transistors. Thus the current in M29, one of the output transistors, depends on the current supplied by M24 (and derived from I1 through M1 and M2) and the channel dimensions of M23, M25, M26 and M29. An identical arrangement is provided for M27 by M19 to M22. With the dimensions shown the idle current amounts to 70uA.

In this circuit we also have two higher-performance current mirrors: M15 to M18 (see figure 3-24) to get the maximum open-loop gain and M5 to M8 (see figure 3-25) to get the highest possible common-mode rejection (now 98dB, up from 94dB in figure 8-16).

The output stage is capable of supplying 1mA peak and can get within 100mV of the rails with a 5kOhm load. This performance can be increased by making M27 and M29 wider.

Though capable of a much higher current without wasting idle power, this rail-to-rail output has the same weaknesses as the previous one: it is very sensitive to capacitive load and the open-loop gain is fundamentally related to the load impedance. With no load it is 105dB; with a 100kOhm load the open-loop gain drops to 101dB, with 10kOhm to 88dB and 1kOhm it reaches a paltry 68dB. This, alas, is an unpleasant fact with rail-to-rail outputs.

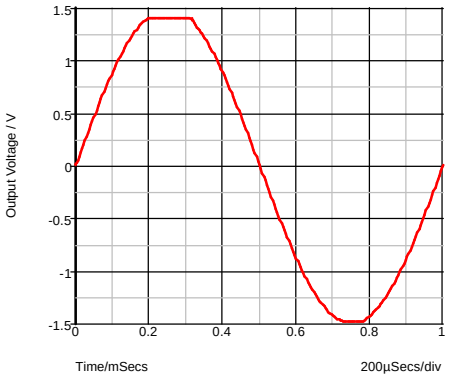


Fig 8-20: Output limits with 5kOhm load.

Larger input transistors are also used in this circuit, which reduces the white noise to 23nV/rtHz (28nV/rtHz in the previous circuit).

In the next circuit (figure 8-21) the polarity of the input stage is reversed and the current mirror for the second stage (M11 to M14) is designed to have the highest possible output impedance, resulting in an increased loop gain. Also note that the primary current has been (arbitrarily) reduced to 5uA.

The lower operating current level has only a minor effect on the sizes of most devices; they still need to be large to obtain satisfactory matching, much larger than the process (0.35u, or the higher-voltage portion of a 0.18u process) would allow. The idle current of the output stage is now reduced to 10uA.

Open-loop gain is 107dB at low frequency and with no load. Capacitive loading is still a problem (but much reduced if the minimum closed-loop gain is higher than 1) and, as before, the closed loop gain is a function of load impedance.

The input operating range now extends from about +0.8V to 150mV below the negative rail. If a single supply is used the inputs can function at or below ground level.

Because of the large dimensions used for the input transistors the white noise level is a relatively low 21nV/rtHz. Note, however, that they are run at twice the level of I1.

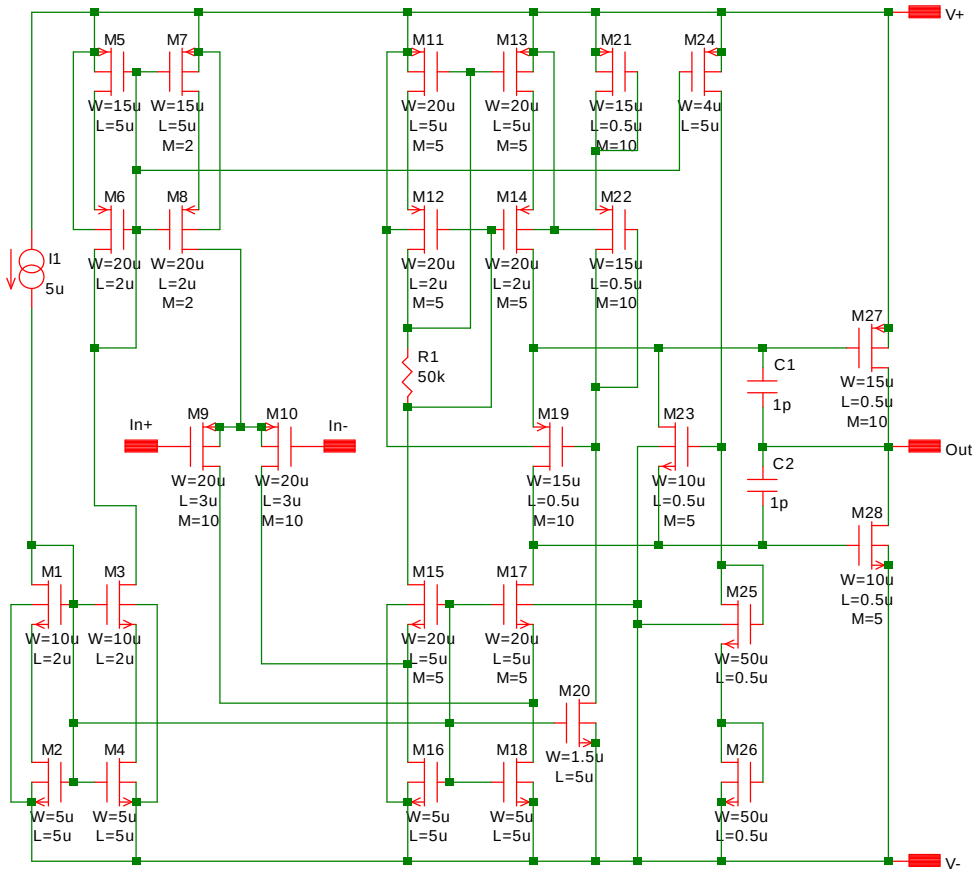


Fig. 8-21: Op-amp with P-channel input.

Now let's extend the operating range of the input by using both p-channel and n-channel devices. In the circuit of figure 8-22 we are adding an n-channel differential pair which takes over when the DC level at the inputs reach about +0.8Volts and the p-channel devices get cut off. There are three voltage regions for the input now: within 0.8 Volts of the negative rail only the p-channel devices are active; from about -0.8 Volts to +0.8V at the inputs, both pairs amplify and within 0.8V of the positive rail only the N-channel devices amplify.

When both pairs are active, the open-loop gain is at a maximum, reaching 120dB. When the common-mode level is either high or low, the

loop gain drops by 10dB. Many schemes have been offered in the literature which hold this gain more constant (e.g. by allowing only one pair to operate at a time), adding another dozen devices. For most applications the benefits of this measure are limited; in fact simultaneous operation of both pairs increases performance (noise, for example drops to 20nV/rtHz, compared to 30nV/rtHz when only one pair is amplifying).

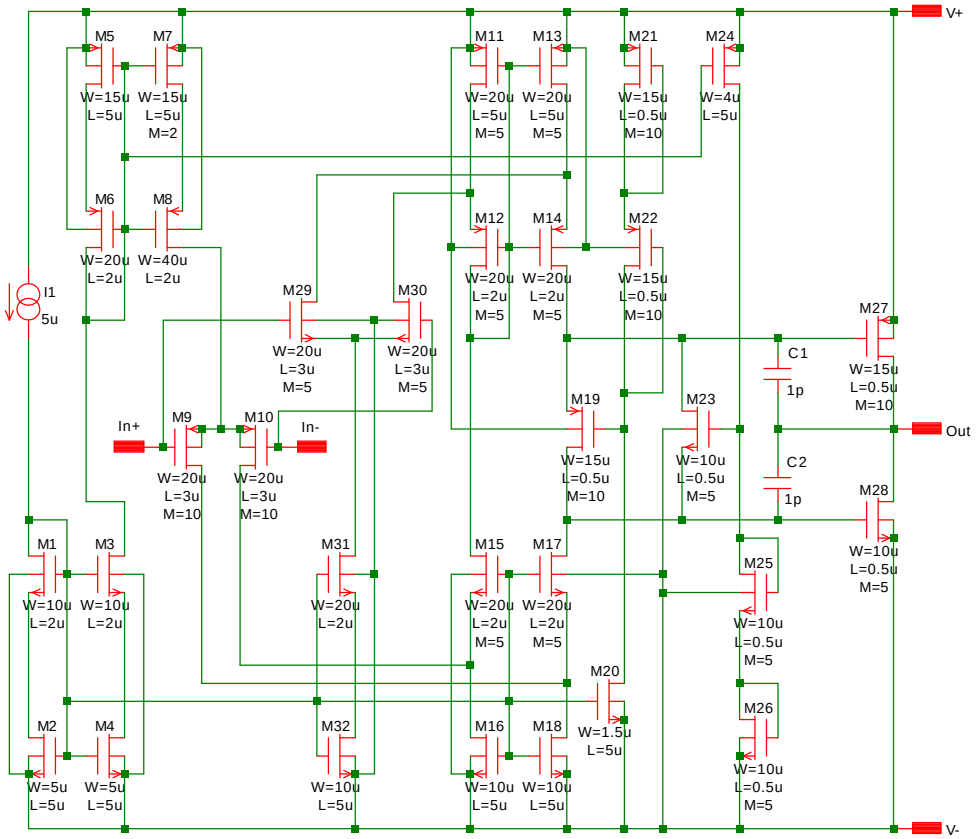


Fig. 8-22: Op-amp with rail-to-rail inputs and output.

The two input stages work with the second stage (M11 to M18) as folded cascodes. The lower part of the second stage (M15 to M18) is a current mirror derived from M1 to M4, set here at about 10uA; the upper part (M11 to M14) mirrors the current again, so that at M19/M23 the currents cancel if there is no input signal. M19 and M23 set the bias current of the output transistors (M27, M28)

With an input signal one or both input stages change the currents in the second stage, resulting in a net positive or negative current through M19/M23, which is translated into a larger current at the output.

A common problem in op-amps using two separate input stages is created by the random nature of the offset voltage. Suppose one pair has an offset voltage of +5mV the other -5mV. As the signal moves from one stage to the other, this causes a jump of 10mV, creating distortion.

Auto-Zero Op-Amps

Auto-zero or **chopper stabilized amplifiers** have been around for decades but continue to evolve. In a modern embodiment two amplifiers are used, checking up on each other.

Each amplifier has a "Trim" input, i.e. a single node which changes its offset voltage in both directions.

A built-in oscillator flips the two switches periodically at a rate of a few hundred to a few thousand Herz. In position A the inputs of amplifier 2 are shorted together and its own offset voltage is amplified (with the open-loop gain) and corrected by feeding the output to the trim input. The required trim voltage is stored in capacitor Ca.

In the second phase of the oscillator the inputs of amplifier 2 are connected in parallel to those of amplifier 1 and its output now feeds the trim input of amplifier 1. With the charge remaining across Ca, amplifier 2 continues to be nulled and thus corrects the offset of amplifier 1. As the oscillator switches back to phase A this correction voltage remains across capacitor Cb. With the high open-loop gains of both amplifiers the offset voltage is now reduced to microvolts. Since the correction is done repeatedly, temperature drift is also much reduced.

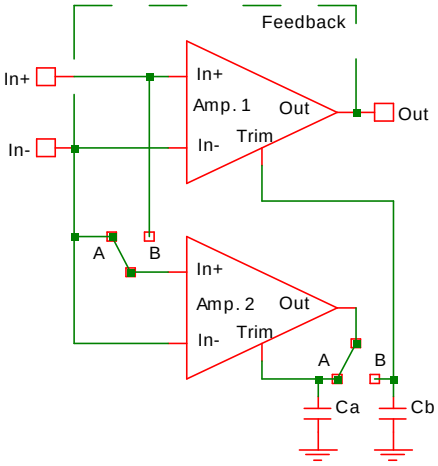


Fig. 8-23: Auto-zero op-amp.

There is an additional benefit. Anything sensed by amplifier 2 *below the switching frequency* is treated as an offset. This includes flicker (1/f) noise, which is completely eliminated. Above the switching frequency the behavior of the auto-zero amplifier is identical to a regular op-amp.

Apart from the higher current consumption because of the additional circuitry there is one drawback: switching noise. At the switching frequency there is a noise peak which also causes (intermodulation) distortion. This effect can be ameliorated by changing the switching frequency at random, i.e. creating a spread spectrum.

Distortion in an Op-Amp

An op-amp is basically a non-linear circuit. The input stage, for example, can accommodate only a small differential voltage before it is limited by the input devices, both bipolar and CMOS.

Feedback reduces the distortion caused by these limitations. Increasing the amount of feedback increases the linearity.

If you increase the open-loop gain by a factor of 10 (20dB), distortion drops by a factor of 10, assuming that, by increasing the gain you have not added more distortion.

The Miller Capacitance

In 1919 John M. Miller was physicist with the National Bureau of Standards when he wrote a paper on how the grid capacitance of a vacuum tube was so much larger in use than measured statically. The voltage gain, he said, multiplies the capacitance between grid and plate. What he described has been known as the Miller effect or the Miller capacitance ever since.

Miller went on to doing research at Atwater Kent, RCA and the Naval Research Laboratory. In 1953 he was awarded the IRE Medal of Honor.

The exact same effect was found in both the bipolar and MOS transistor. In most applications it is detrimental, limiting the frequency response; in IC op-amps, however, it has been helpful, greatly decreasing the size of the compensation capacitance.

John M. Miller: "Dependence of the input impedance of a three-electrode vacuum tube upon the load in the plate circuit", Scientific Papers of the Bureau of Standards, 1920, pp. 367-385.

9 Comparators

To most people a comparator is merely an op-amp without feedback. With the very large open-loop gain the output abruptly traverses the entire available voltage range when one input passes the level of the other.

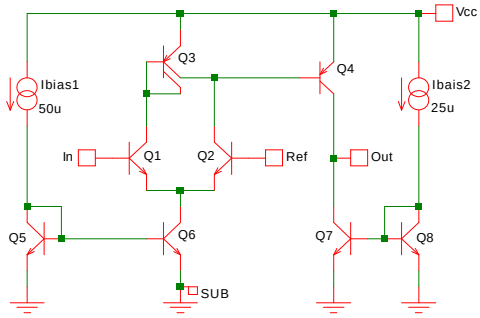


Fig. 9-1: Simple but accurate bipolar comparator.

This is true for the majority of comparators, but there are also some refinements and variations. Let's examine them.

The first circuit is indeed of the common variety: an input differential pair (Q1, Q2), a current mirror active load (Q3) and a second stage (Q4), giving a voltage gain of about 95dB.

The second stage is run at half the current compared to the input stage, so that it switches when the differential pair is in balance. It uses a separate current mirror (Q7, Q8) for a good reason: Q7 saturates. If we were to run Q7 off Q5 (as Q6 is), it would grossly decrease the collector current of Q6 as it saturates.

This comparator, using bipolar transistors, requires a small input current; with an operating current of 50uA (25uA for each input transistor at balance) and a minimum hFE of 100, that amounts to 0.25uA. We could of course decrease the operating current, but at the expense of speed and noise.

Also, the reference voltage (i.e. the common-mode voltage) cannot drop below the VBE of the input transistors (plus the saturation voltage of Q6), otherwise the input stage is simply cut off. At the upper end the common-mode range stops at about 0.2V below Vcc, when the input transistors saturate and cut off Q3 and Q4. On the other hand, Vcc can be as low as 1 Volt.

A simulation for a high-gain circuit like this one is best set up by connecting two voltage sources to the inputs. One is steady DC (say 1.5 Volts) while the other one is swept from 1mV below this value to 1mV above it. You will see the output change drastically very close to the zero difference at the input. There is very little built-in error because Q1, Q2 and both sides of Q3 operate at the same collector-base voltage; there is only a

small second-order error due to the fact that the collector-base voltage of Q4 is larger.

But don't let this observed accuracy fool you into believing that this is what will happen in production. Move on to a Monte Carlo analysis and you will find that the offset voltage of the differential pair moves the switching point (by about ± 1mV, depending on the process and the size of the transistors).

The bipolar design can be directly translated into CMOS, with a logic stage added at the output. The gain is now 110dB.

There is no (DC) input current, but the limitations concerning the common-mode range still apply.

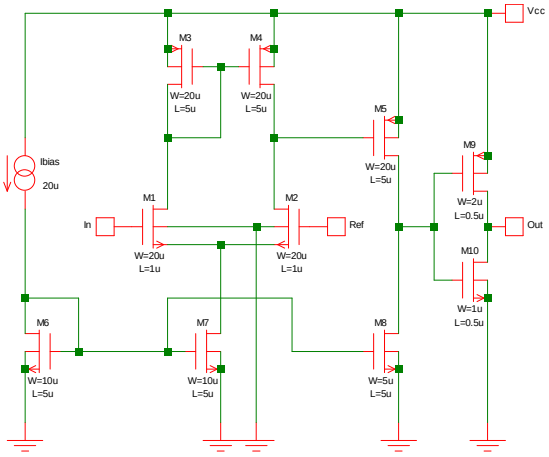


Fig. 9-2: CMOS version of figure 9-1.

In a CMOS circuit saturating current mirrors need not be feared; the current in M8 can be derived from M6. The fact that the drain of M8 can end up very close to ground has no adverse effect on M7. A word about the transistor dimensions: the logic stage at the output is designed for a 0.35u process, all other channel lengths and widths need to be this large even for a process capable of smaller sizes. M1 through M4 require a large area for adequate matching (in fact, offset can be further reduced by increasing their sizes) and the 5um channel lengths reduce dependence on supply.

Quite often **hysteresis** is required in a comparator, i.e. the threshold is higher when the input increases and lower when it decreases. For example, if you have a "low fuel" warning light you don't want this light to flicker on and off as the fuel sloshes in the tank, so you set the threshold to a low level as the fuel is consumed and to a higher level as the tank is filled. In figure 9-3 two features have been added. Replacing the simple current mirror, Q3 and Q4 form a flip-

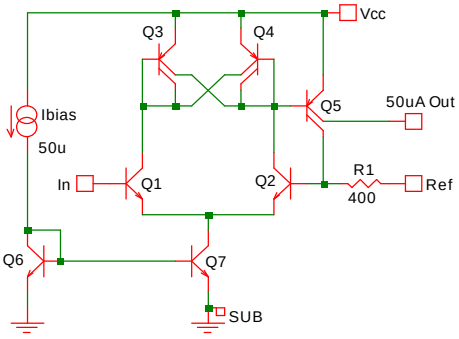


Fig. 9-3: Comparator with hysteresis.

flop with precisely controlled gain, giving the circuit a snap action. In addition, the diode connection of Q4 makes Q5 into a current mirror (each collector sources one-half of the current). This current is fed into a 400 Ohm resistor, causing the reference voltage appearing at the base of Q2 to increase by 10mV (the resistor value can, of course be changed to increase or decrease this value). The voltage at Ref must be capable of sinking the collector current of Q5.

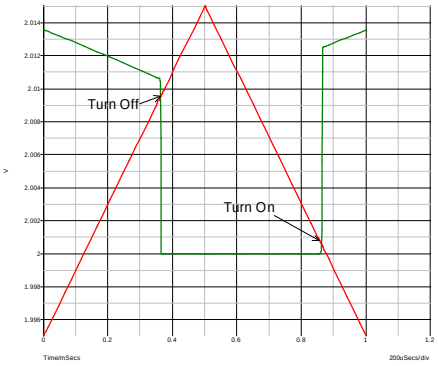


Fig. 9-4: Switching levels with hysteresis.

As the input voltage decreases from some value above the reference voltage, the current out of the output terminal abruptly increases from zero to about 25uA (the exact level depends on the output voltage because of the Early effect). You now have to increase the input voltage to 10mV above the reference level to turn the output current off.

Figure 9-5 shows the same circuit in CMOS, with the current output (M8) opposed by a current sink of half the level (M13), and a logic stage

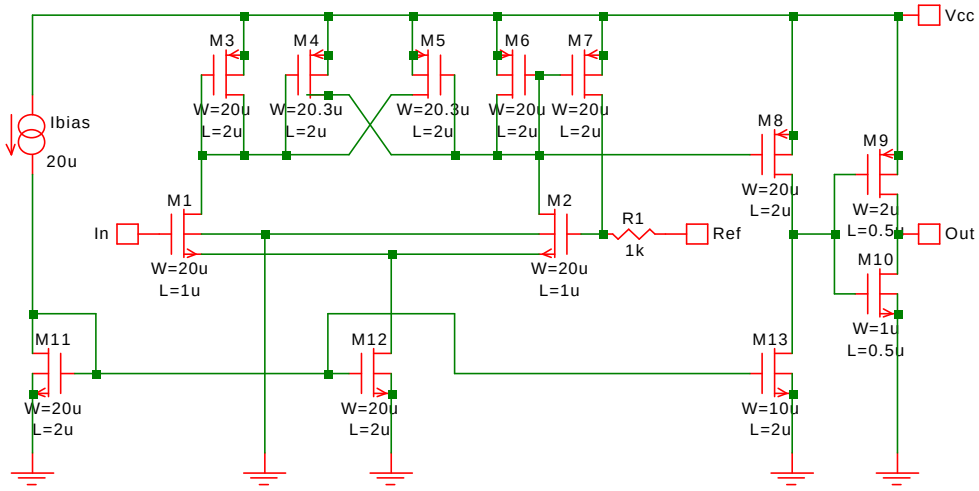


Fig. 9-5: CMOS comparator with hysteresis.

added. CMOS has an advantage here in that the custom sizing of the transistors allows the amount of positive feedback to be set in precise increments (M3-M6). Note that the operating current (Ibias) has been

reduced and the value of R1 increased, resulting again in a hysteresis of 10mV.

Ibias will most likely be derived from a resistor value (and, perhaps, a bandgap reference voltage). R1 will track this resistor value and thus the hysteresis is remarkably accurate and stable with temperature.

A comparator with hysteresis requires some thought before simulating or testing. The two different thresholds have to be approached in the proper sequence. In a simulation this can be done with a transient analysis, i.e. letting the input voltage increase until it exceeds the upper threshold, then decreasing it until a level below the lower threshold is reached. Similarly, in testing the input is ramped up until switching occurs, and then ramped down until the output changes states again.

In the examples so far NPN or N-channel transistors have been used for the input differential pair. This is a disadvantage when input signal and reference are near ground level, unless you have a split power supply.

By converting the input to PNP or p-channel and a couple of design refinements, a comparator can be made to work *at ground level*, even if there is only a single supply.

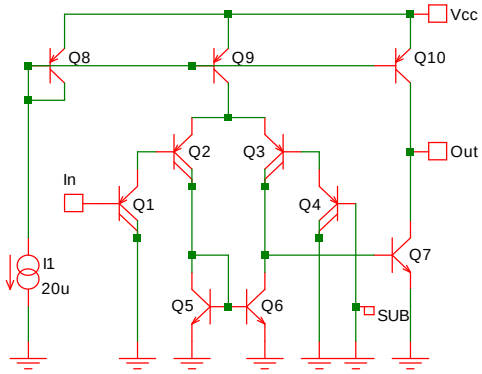


Fig. 9-6: A Darlington input allows the inputs to be at ground level.

In figure 9-6 a Darlington input stage is used not to decrease the input current, but to allow the comparator to operate even if the input drops slightly below ground. With as much as 400mV below ground at the input Q1 is still in its active region. At that point the base of Q2 is about 200mV above ground (at room temperature), which is sufficient to keep Q5 from being cut off. (Strictly speaking the input stage is not a pure Darlington connection, since the collectors go to ground).

This circuit has a definite upper temperature limit (about 100^{o}C) and is rather slow because there is no discharge path at the bases of Q2 and Q3. Since the primary object is not a low input current, however, there is no reason why we could not place two additional small current sources (like Q9) at these points.

Though rarely needed in an ASIC, the ideal comparator has a rail-to-rail input (it already has a rail-to-rail output).

This is actually quite easy to achieve: two input differential pairs, one n-channel, the other p-channel; mirror the currents of one and sum the result with the currents of the other.

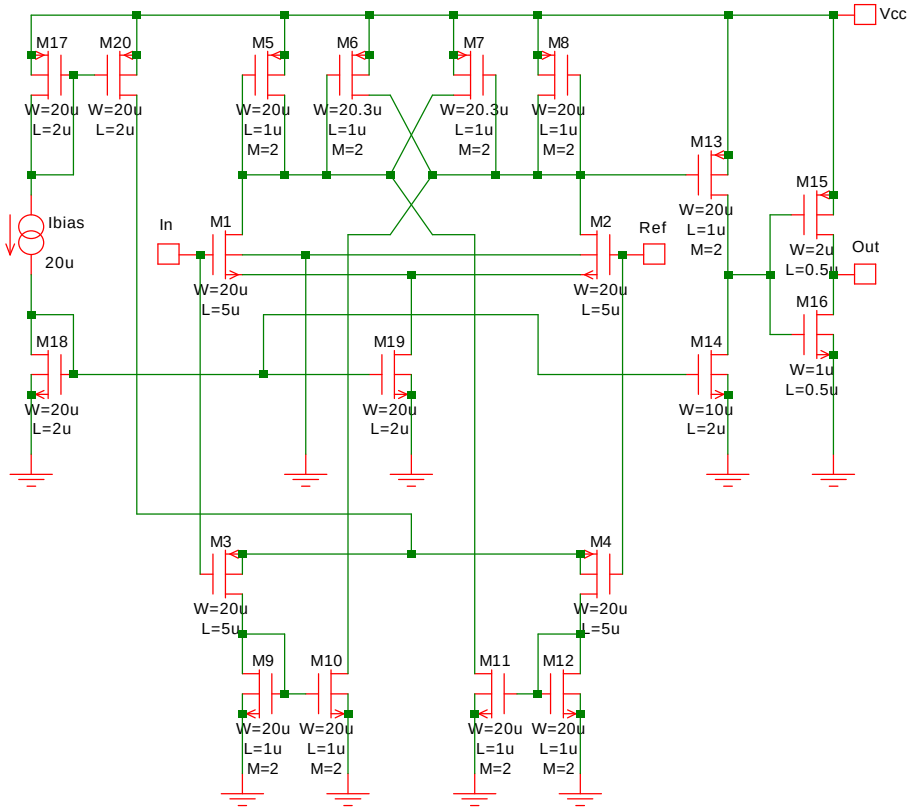


Fig. 9-7: CMOS comparator with rail-to-rail inputs

In this example the active load of figure 9-5 was chosen, again with sufficient positive feedback to give a snap-action (M5-M8). Note that M5-M8 and M9-M12 have a considerably large w/l ratio compared to the corresponding M1 - M2 and M3 - M4 to allow the input to go slightly beyond Vcc and ground.

Current Comparators

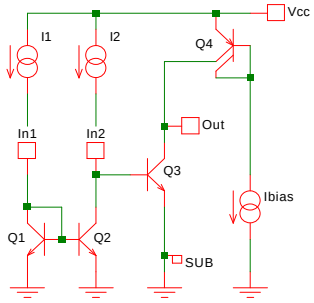


Fig. 9-8: Bipolar Current comparator.

When you use the word comparator you automatically assume that voltages are compared. But this does not always have to be the case, sometimes it is useful to compare currents.

With a simple current mirror (Q1, Q2) even a small difference in the magnitudes of I1 and I2 will show up quite drastically at the base of Q3, turning it on or off.

The base-current error is eliminated if Ibias is set at twice the level of I1 and I2. The only remaining error is due to the Early effect of Q3, which is easily reduced by adding another NPN stage and using a more sophisticated current mirror in place of Q4.

The CMOS version is almost identical. There is of course no base-current error, but the comments above about the Early effect (or channel-shortening) apply. Yet even without any improvements both circuits switch abruptly within 0.0006% over a wide temperature range. Matching variations, however, are another matter and you may have to make the input current mirrors quite large to get enough accuracy. Find out with a Monte Carlo analysis.

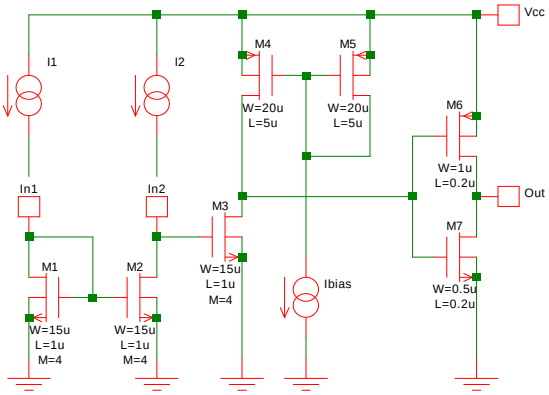


Fig. 9-9: CMOS version of current comparator.

10 Transconductance Amplifiers

For a while it looked like there would be a second universal building block, and the concept was called the *Operational* Transconductance Amplifier. But the OTA has rather severe limitations and there is no danger that it might de-throne the op-amp anytime soon.

Let's examine the concept using a simple bipolar design. Just as in an op-amp there is a differential input pair (Q1, Q2). Its collector currents are mirrored separately by Q3-Q5 and Q6-Q8. One of the mirrored currents goes directly to the output, the second one is mirrored again (by Q9-Q12) and then opposes the first one at the output. No matter what value is chosen for the operating current (Igain), the two currents at the output have the same value (without an input signal) and the output voltage is at ground.

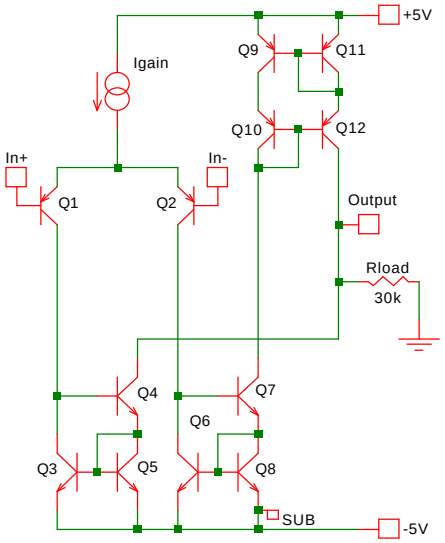


Fig. 10-1: A simple bipolar transconductance amplifier.

Ignore Rload for a minute. As we have seen in chapter 4 a bipolar transistor has an emitter resistance

$$r_e = \frac{k*T}{q*Ie}$$

where k and q are constants, T is the temperature in Kelvin and Ie is the emitter current. The term for r_e is *dynamic* emitter resistance (often called "little r_e") because it changes with emitter current. At Ie = 1mA it amounts to roughly 26 Ohms (at room temperature), at 100uA 260 Ohms, at 10uA 2.6kOhm, i.e. it is inversely proportional to Ie. (There is also a constant resistance in series with r_e, the physical resistance between the emitter contact and the base-emitter junction; this becomes significant at higher currents).

The transconductance of a bipolar transistor is simply:

$$g_m = \frac{1}{r_e}$$

Thus with an emitter current of 100uA the transconductance is (1/260) Ohms, i.e. a 1mV signal at the base cases a change in collector current of 3.8uA. In a differential stage the transconductance is half of that since there is an r_e in each transistor (and we double the total current so that each emitter receives 100uA).

In figure 10-1 the currents are mirrored in a ratio of 1:1 so that the collector currents of Q1 and Q2 appear unchanged at the output. With no signal at the input they cancel each other but, as one input is moved up or down, one current becomes larger and the other one smaller by the same amount. Thus the total transconductance is doubled and we have the same value as for a single transistor.

Without some DC resistance at the output a transconductance amplifier is really quite impractical. Even the slightest mismatch in any of the transistors would slam the output voltage into one of the supply rails; we need some impedance like Rload to keep this voltage near the center. Rload converts the current output into a voltage output, which means that we no longer have a transconductance amplifier but simply a voltage amplifier (with a high output impedance to boot). Very few of the OTAs are actually used as transconductance amplifiers.

With Rload back in the circuit the total voltage gain is now simply:

$$A_v = \frac{Rload}{r_e} = \left(\frac{q}{k*T}\right)* Ie* Rload$$

So we have an amplifier whose gain can be varied (over a wide range) by varying a current .

And herein lies the problem. *The input signal also varies the current*, and thus the gain changes with the amplitude of the signal. The result: distortion. With a small signal at the input this may be tolerable for some application, but not with a large signal. Here is the tally:

Input Signal		Igain=1uA	Igain=10uA	Igain=100uA
		Gain		
		-5dB	14dB	32dB
		Distortion		
10mVp		0.3%	0.2%	0.1%
20mVp		1.2%	0.9%	0.3%
50mVp		6.2%	5.1%	1.6%
100mVp		16%	15%	8%

So, if you have to handle a signal greater than about 20mVp, this circuit is a poor choice. You cannot use feedback or emitter resistors (for Q1 and Q2) to linearize it, it would interfere with the variable gain.

There is another problem, not just for this circuit but for all such schemes: offset. A mismatch in not only the input stage but all three current mirrors will show up as an offset (and added distortion) at the output, increasing in magnitude as Igain increases. In this circuit this amounts to 60mV worst-case at 100uA. Also, remember that bipolar transistors have input currents.

There is help, though, and as usual you need to add a few more devices. If we connect diodes to the inputs and feed the signal in through a resistor, we have something very similar to a current mirror (e.g. figure 3-1). The input impedance is low because of Q16 (at 20uA r_e amounts to 1.3kOhm at room temperature) and R22 converts the input voltage into a current. A 500mVp input signal causes so little change in voltage at the base of Q1 that the distortion is down to 0.3% at 1uA and 0.01% at 100uA. The offset voltage still persists, amounting to 60mV again worst case at 100uA.

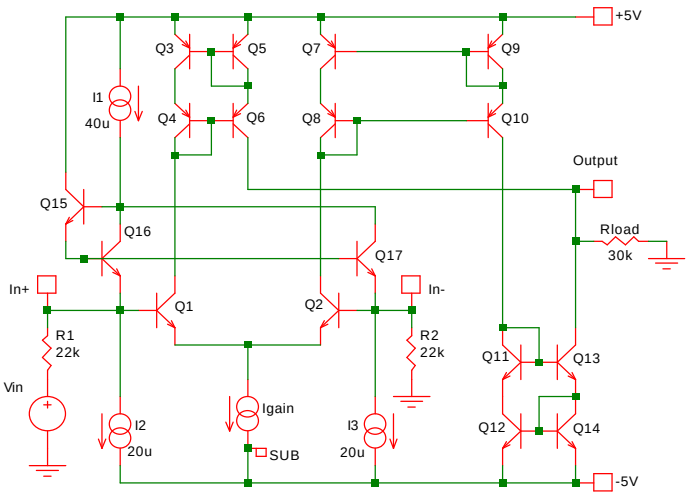


Fig. 10-2: An improved version of the bipolar transconductance amplifier with linearizing diodes.

Both sides of the input pair need to be treated equally, including the addition of the dummy resistor R2 to avoid worse offset problems. The current mirrors used here are of the highest precision, sacrificing low operating voltage for accuracy.

Q15 aids to remove the base currents for Q16 and Q17 from I1, but even with this measure the ratio between I1 and I2/I3 needs to be precise; any mismatch will increase the offset voltage and the input current (250nA max. with perfect matching).

The one great feature of a circuit like this is the precise control of gain over a wide range. Figure 10-3 shows the gain (in dB) vs. Igain. A linear change in current results in a logarithmic change of gain. Thus, for audio applications, you can control the volume (a logarithmic function) with a linear current (or voltage). The accuracy is within ±0.2dB. True to the exponential nature of the base-emitter diode, the gain changes 20dB per decade of current.

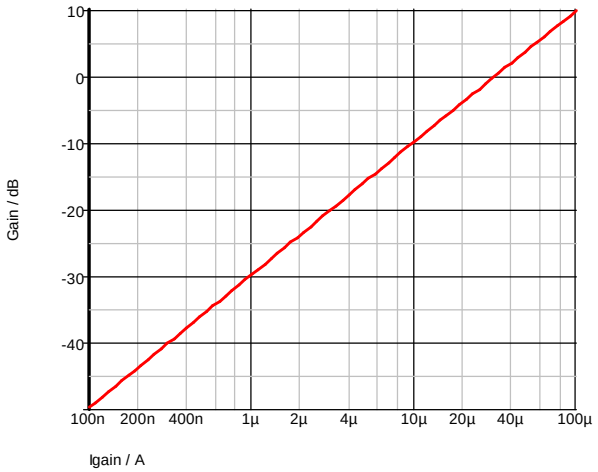


Fig. 10-3: Gain is a precise logarithmic function of Igain.

Because the diode-connected transistors (Q16, Q17) track the input transistors, gain is virtually unaffected by temperature. If Igain is derived from a resistor made from the same layer as R1, R2 and Rload, the gain is also unaffected by absolute variations.

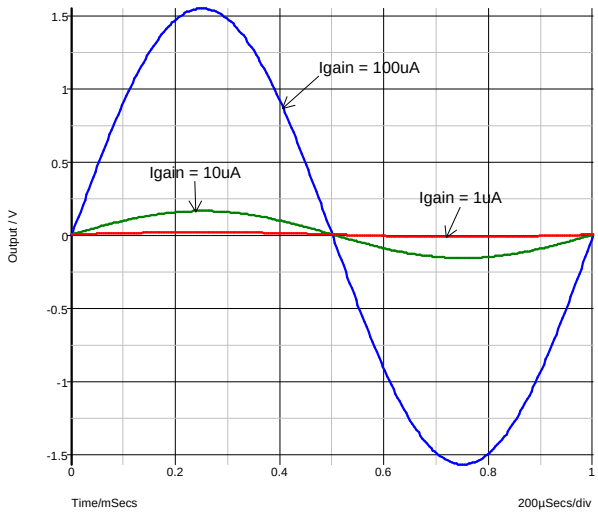


Fig. 10-4: Output waveforms (1kHz).

Figure 10-4 shows the waveform that appears at the output. There is a very large change in the level of the signal, which is of course the purpose of the circuit. Because of the offset voltage a "transconductance" amplifier is best suited for audio and filter applications with the output capacitively coupled to the next stage.

The concept works in CMOS too, but the fundamentals are different. A CMOS transistor naturally takes a voltage at the gate and delivers a current at the drain, and this transconductance varies as the square of the operating current.

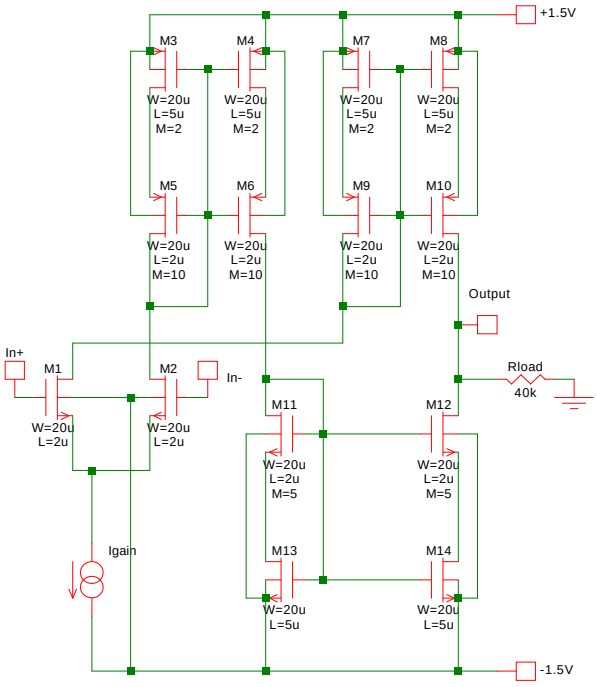


Fig. 10-5: A CMOS equivalent of figure 10-1.

Figure 10-5 is the same configuration as figure 10-1, with NPN transistors replaced by N-channel devices and PNP transistors by p-channel ones. The devices are quite large (M5, for example, has a total width of 200um with the multiplier M set at 10) and, as we will discover shortly, the sizes chosen are still marginal.

In this and the next example a dual power supply of ± 1.5 Volts is used. This may be an impractical value for you, the choice was made to simplify the discussion of input and output DC levels. In the real world you may be forced to use a single 3-Volt (or 3.3-Volt) supply, in which case the inputs and the output have to be biased at half the supply voltage.

This circuit uses a 0.35um process, necessary because the high-accuracy current mirrors cannot tolerate an output voltage of less than about 0.6 Volts across them. If you were to use a process with smaller dimensions you would have to reduce each mirror from four to two devices and pay the penalty of much reduced accuracy.

A CMOS transconductance amplifier suffers from the same non-linearity as a bipolar one. Distortion is tolerable only for small input signals. With a ±40mVp input it amounts to 0.1% at 100uA, 0.7% at 10uA and 1.4% at 1uA. When the signal is increased to ±75mVp (which results in the maximum output swing possible, ± 0.9V) the distortion increases to 0.8% at 100uA, 2.3% at 10uA and 4.5% at 1uA.

In the gain vs. operating current plot (figure 10-6) the range is extended down to 1nA just to show the wide range achievable. You notice, however, that at the high-end the circuit deviates from a pure logarithmic behavior; to straighten this line the transistors would have to be even larger.

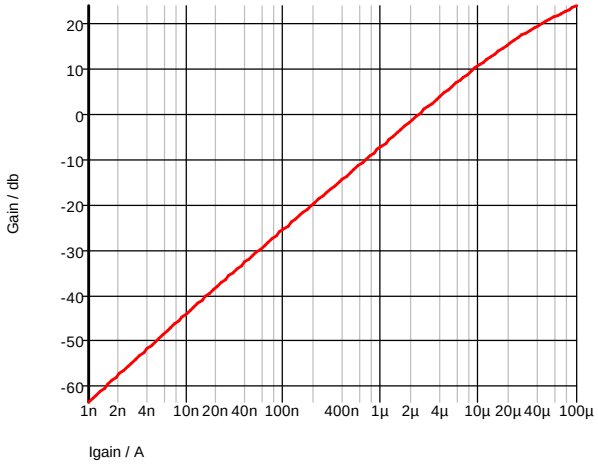


Fig. 10-6: Gain (in dB) vs. Igain.

Transconductance of an MOS transistor is temperature dependent and thus there is a decrease of gain (at any current) of about 2dB from 0 to 100^{o}C. Also, a CMOS transconductance amplifier has the same offset problem as a bipolar one; at 100uA this amounts to ± 30mV. Unlike the bipolar version, however, this circuit has no DC input current.

A rather complex scheme has been developed to linearize the input stage and still have gain control (see references). To do this 15 more transistors are needed (M15 to M29, figure 10-7).

M17 is the key device. This "diode-connected" transistor is the same size as M1 and M2, the input differential pair, and all three devices share a mirrored Igain current (M23) at their sources. The drain/base node of M17 receives the same amount of current from M24 through the current mirror M26 to M29. M25, a cascode transistor, has been added to improve the matching of the mirrored currents.

The current into the drain/base node of M17 is also shared by M16 and M18, but their current is governed by the fact that they each are part of another differential pair (M16/M19 and M18/M15) whose operating currents are set at (3/4)Igain. M15 and M19 are twice as wide as the other five devices in the input row.

This complicated use of ratios serves to extend the range of input voltage over which the differential pair is linear. The optimum is reached with a ratio (between M15/M18 and M19/M16) of 2.155, exceeding ± 1 Volt. For our case here this is of little consequence, 75mV causes an output swing of ± 0.9 Volts, the maximum the circuit can handle.

At this level the distortion now amounts to 0.1% at 50uA, 0.2% at 5uA and 0.3% at 0.5uA.

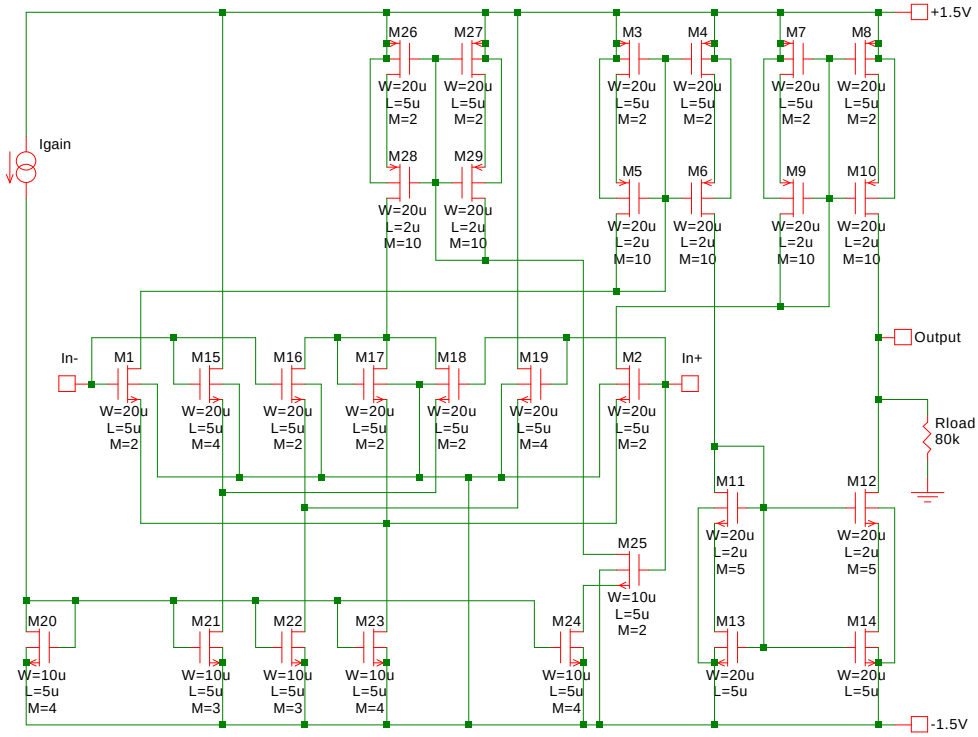


Fig. 10-7: Transconductance amplifier with linearized input.

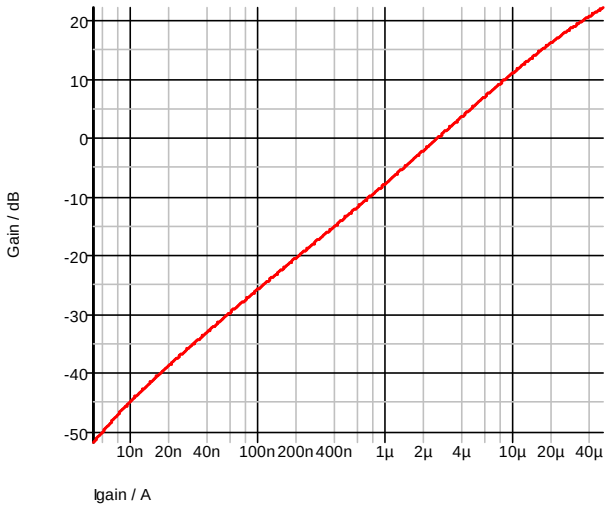


Fig. 10-8: Gain (in dB) vs. operating current (Igain).

With the device dimensions shown the circuit cannot handle much more than 50uA (Igain) and starts deviating slightly from the ideal logarithmic line above 20uA. Because of the many additional devices there is also a slight deviation below about 10nA. Even so, the gain control has a range of more than 70dB.

Note that the output impedance is

rather high; a buffer may be needed.

The problem with the offset voltage is still present, only slightly reduced by the lower gain. Figure on a ± 20mV uncertainty at the output. For this reason transconductance amplifiers are primarily used in audio and filter applications where the output can be capacitively coupled.

11 Timers and Oscillators

Summer of 1970. The economy was at the bottom of the cycle and Signetics, the promising young company I had joined just two years before, laid off half of its employees.

Disgusted with the turn of events, I decided it was time to strike out on my own and rented space between two Chinese restaurants in downtown Sunnyvale, California. Signetics (now Philips) lent me the equipment I needed and gave me a one-year contract to develop a new IC.

The idea for the new IC came from the work I did at Signetics on the phase-locked loop. I had needed an oscillator whose frequency could be set by an external resistor and a capacitor and was not affected by changes in either supply voltage or temperature. Several products resulted from the basic design, among them the NE566 Voltage-Controlled Oscillator.

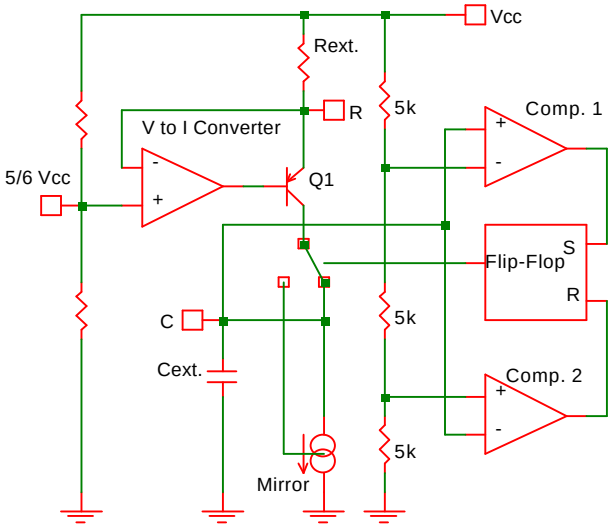


Fig. 11-1: The basic 566 Oscillator. In the actual circuit the comparators and the flip flop are combined in one Schmitt trigger.

The oscillator contained first of all a voltage-to-current converter. The reference voltage at the positive input terminal of the op-amp is not regulated, it is simply a fraction of the supply voltage. Feedback to the op-amp keeps the voltage across the external resistor at the same level and thus the current through the resistor becomes (1/6*Vcc)/Rext.

Depending on the state of the switch controlled by the flip-flop, the external capacitor is either charged with the current, or discharged with a current of the same magnitude through a 1:1 current mirror.

There is a divider with three identical resistors, producing 1/3 Vcc and 2/3 Vcc at the two taps. Two comparators, referenced to these taps, watch the voltage across the external capacitor. If it moves above 2/3 Vcc, comparator 1 sets the flip-flop, the switch diverts the current to the mirror and the capacitor is discharged. When the voltage across the capacitor reaches 1/3 Vcc, comparator 2 resets the flip flop and the capacitor is charged again.

This endless cycle produces a triangle-wave. The amplitude is dependent on the supply voltage, but so are the charge and discharge currents and *the two effects cancel each other*. Except for small errors inside the IC, such as the offset voltages of the op-amp and comparators and the matching in the current mirror, the frequency is exactly:

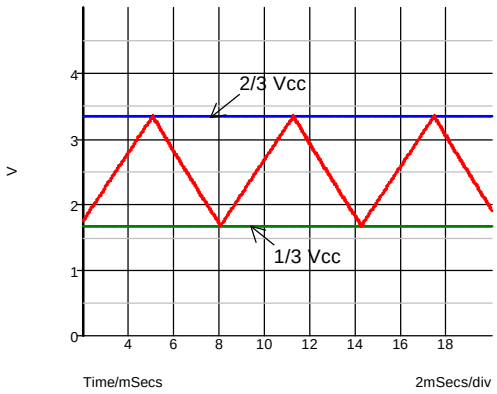


Fig. 11-2: Triangle-wave produced by the 566 oscillator.

$$f = \frac{1}{3 * R * C}$$

What I proposed to Signetics was this circuit, modified so it could also be triggered and produce a single cycle only, i.e. it would be both an oscillator and a timer.

The project almost didn't get off the ground; the engineering staff didn't think much of the idea. Timers at the time were put together from an op-amp or comparator and a few discrete components, including a Zener diode or two. They argued that such a design would cut into the sales of their present ICs. But the marketing manager, Art Fury, over-rode them; a man with immense practical experience, he simply had the gut feeling that such a timer would sell.

It was a one-year contract and designing the circuit took half of that. No computer analysis then, the circuit had to be laboriously breadboarded. When everything was working I wrote a development report and gave a design review at Signetics. The design passed without any comments.

But something wasn't quite right. I felt that I had missed something, that I could do better. It bothered me that the design required nine pins, which was about the most unfortunate number I could have picked. There was an 8-pin package; the next higher number was 14.

I started on the layout. In 1971 this meant sitting at a drawing board for several weeks, fitting devices together into the minimum rectangular shape and checking each dimension by hand. Then, about two weeks after the design review, on the way home after work, it suddenly hit me: what would happen if I got rid of the voltage-to-current converter and charged (and discharged) the capacitor directly with a resistor? That would bring the pin-count down to eight.

I made a U-turn, went back to work and tried it. Sure enough, the timing didn't change as I varied the supply voltage. It was my own limitation that had made me assume that only a *linear* relationship between charge current and end-voltage would cause the cancellation effect. Even though the charging of a capacitor through a resistor causes an exponential rise of the voltage, the cancellation was just as effective. In fact, having eliminated the voltage-to-current converter, I now had not only a smaller but also a more accurate circuit.

I made the changes in the circuit but didn't bother to request a second design review. I only told Art Fury, who was pleased; an 8-pin package was then significantly less expensive than a 14-pin one.

It took another five months to draw the layout, cut the patterns on Rubylith (by hand), spend endless hours hunched over a light-table to check dimensions and connections (again: by hand, no computers), make a mask and a prototype wafer and then evaluate the IC, which Art Fury decided to call the NE555.

In the meantime, one of my former colleagues left Signetics to join a start-up. The first circuit this start-up brought to market was the timer I had described in my design review. Time-wise they beat Signetics by two months, but when the real 555 came out, they had to withdraw their version very quickly.

The market reaction to the 555 timer was truly amazing. Art Fury made history by bringing out the circuit at an unprecedented low price, 75 cents. I had deliberately made the design flexible, but nine out of ten applications were in areas and ways I had never contemplated. For months I was inundated by phone calls from engineers who had a new idea for using the timer. To this day the 555 has been the best-selling IC every year, copied by numerous companies. Except for a CMOS version, the design has never been changed.

Looking at the design now, 33 years later, there are many areas where it can be improved with the design techniques we have learned since and with the enormous benefit of computer simulation. So, let's look at the actual 555 timer and then a version which benefits from 33 years of progress.

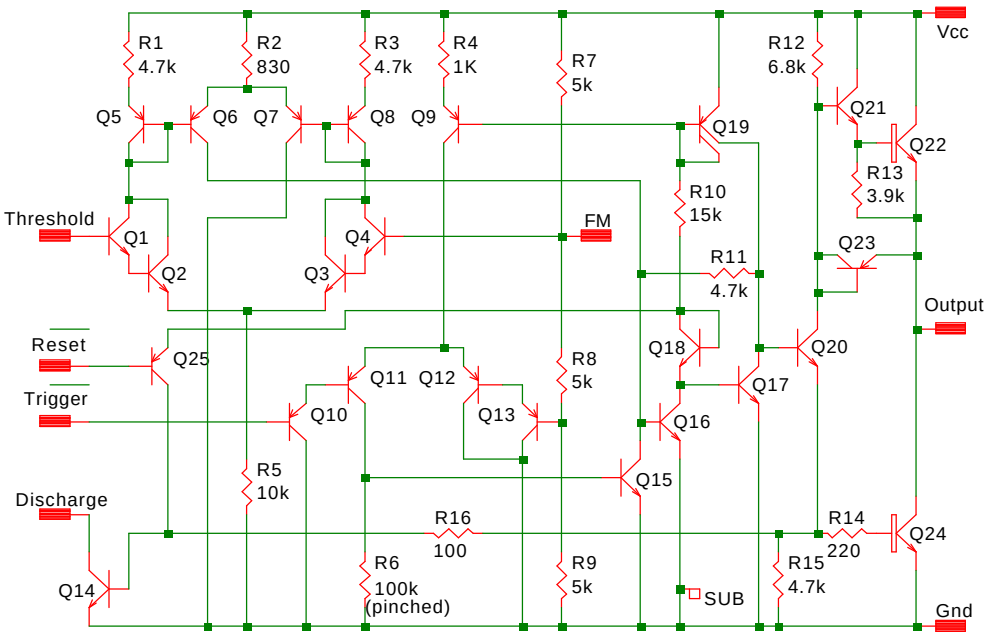


Fig. 11-3 The original 555 timer.

Both comparators use Darlington input stages. This makes the timer fairly slow, but allows an extreme range of external resistance. Comparator 1 consists of Q1 to Q8. The four PNP transistors form a current mirror with gain, provided by the unequal emitter resistors.

The output of this comparator feeds into a 4.7kOhm resistor (R11), which is part of the cross-connection in the flip-flop (Q16, Q17). Comparator 2 (Q10 to Q15) resets the flip-flop.

The output stage, which must be able to sink or source some 200mA, is controlled by Q20. In the high state the Darlington pair Q21/Q22 delivers the current, but at a cost of a voltage drop of about 2 Volts. In the low state Q24 receives sufficient base current to work alone up to about 50mA; beyond that, as the voltage drop increases, Q23 feeds extra current into the base circuit.

There are several flaws in this design, indicative of the early period of IC design (and the inexperience of a rookie designer). Neither comparator is well balanced, showing offsets of as much as 30mV. The circuit can get away with that because the voltage swing is quite large.

The operating currents are quite large; the lateral PNP transistors run at up to 1mA. That was acceptable at the time since the devices had 10um geometries; today it would be excessive.

The output stage consumes a considerable amount of current in the low state and, during switching, both output transistors are on for a brief period of time, producing a current spike in the supply.

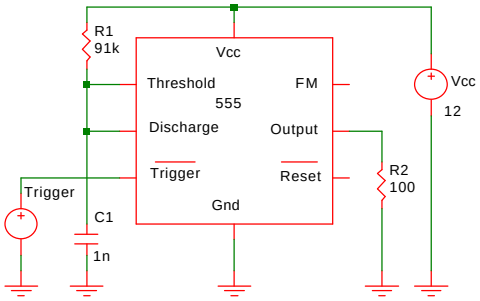


Fig. 11-4: Timer connection of the 555.

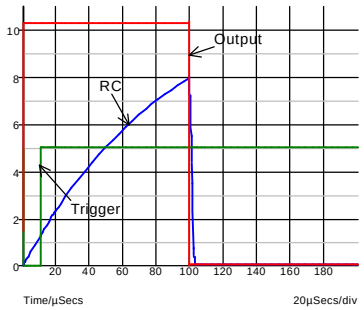


Fig. 11-5: Timer waveforms.

In the timer connection the period starts with a negative-going trigger pulse, which resets the flip-flop through comparator 2 and moves the output high. When the voltage across C1 reaches 2/3 Vcc, comparator 1 sets the flip-flop, C1 is rapidly discharged and the output moves low. Despite the bad offset voltage the accuracy is quite remarkable: the error in timing is around 1% with a temperature coefficient of 24ppm/^{o}C. The timing formula is:

$$t = 1.1 * R1 * C1$$

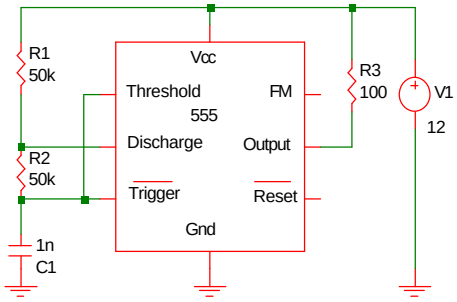


Fig. 11-6: Oscillator connection of the 555.

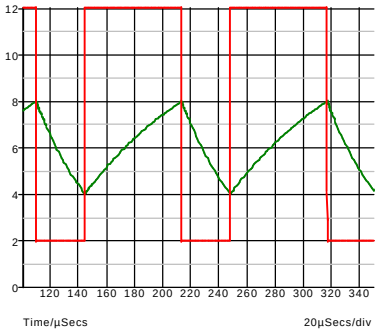


Fig. 11-7: Oscillator waveforms.

In the oscillator connection there are two external resistors and the voltage across C1 moves between 1/3 Vcc and 2/3 Vcc with a frequency and duty cycle of:

$$f = \frac{1.46}{(R1 + 2R2)C1} \qquad DutyCycle = \frac{R2}{R1 + 2R2}$$

It is not quite possible to achieve a 50% duty-cycle; charging and discharging the timing capacitor through a resistor connected to the output is not such a good idea; the high and low voltage drops are unequal and have significant temperature coefficients.

There is a CMOS version of the 555 and a redesign for operation from a single battery cell (see references), but the circuit is still being sold today in its original form, despite the fact that much better performance is possible with more modern design techniques. Here is my candidate:

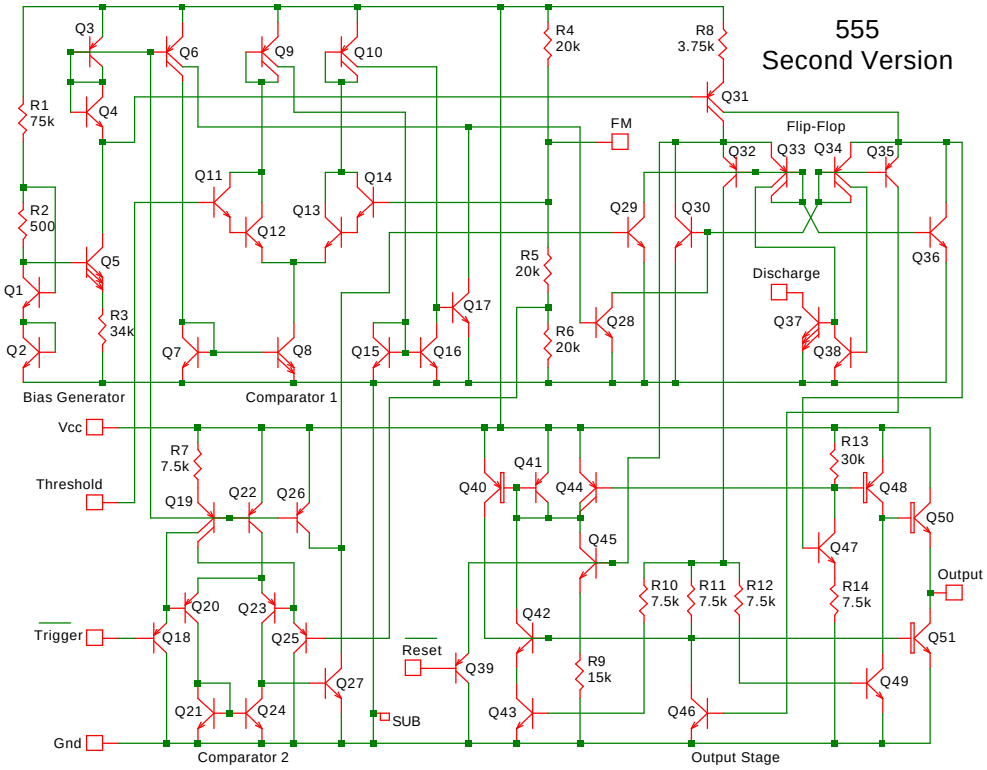


Fig. 11-8: An improved version of the 555 timer, 33 years after the original design.

First off, the new timer gets a proper bias circuit (Q1 to Q5) to hold the operating currents more constant over the wide supply voltage range. This (and a few other steps) extends the operating voltage down to 3 Volts.

Comparator 1 (Q6 to Q17) now has a balanced active load (Q15, Q16) which reduces the error in the timer mode to about 0.5% and the temperature drift to 3 ppm/^{o}C without any loss in speed. The change in timing from 3 to 15 Volts is a mere 0.05%.

There are two changes in comparator 2 (Q18 to Q27): a small operating current for the outer Darlington transistors, which greatly

improves switching speed, and a balanced active load, which makes the trigger level considerably more accurate.

The flip-flop (Q28 to Q36) is a new design; it operates in a current-mode for maximum speed at the lowest possible current. The two 50uA currents generated by Q31 are split by a pair of lateral PNP transistors; one quarter of the current is fed into the base of the opposite flip-flop transistor, another quarter turns the reset transistor on and off and one half of the current is used to steer the output stage. The voltage swing at the collectors of the flip-flop transistors (Q30, Q36) is 2VBE.

The most significant change is in the output stage. The base current for the lower output transistor (Q51) is no longer derived from a resistor. A small amount of current is injected into the bases of three transistors, forced to be equal by the three resistors R10, R11 and R12. This (plus an additional current delivered by Q45) starts a *positive feedback loop* formed by Q40, Q41 and Q42. Q40 is about seven times the size of Q41 and Q42 has one emitter while the output transistor has 24. This loop then provides whatever current is needed to keep Q51 fully turned on.

Positive feedback loops are always dangerous, they can run away or refuse to turn off. In this case the loop is contained by the collector resistance of Q43 and can be opened up by turning Q43 off.

Replacing the Darlington configuration in the upper part of the output stage with a compound (PNP/NPN) transistor reduces the voltage drop. Base current for this part is provided by Q47. Q44, Q46 and Q49 aid in turning the power devices off rapidly and eliminate the large transient current.

With these measures the current consumption is now down to 0.85mA from 3mA (typical) at 5 Volts. At 15 Volts the circuit consumes 1.2mA (down from 10mA). Minimum operating voltage is 2.5 Volts (-40^{o}C to 100^{o}C).

Shortly after the 555 came out Intersil announced a CMOS version. It was (and still is) done in a 15-Volt process, which requires large dimensions and is inherently slow. The circuit is not directly compatible with the bipolar version, lacking high current outputs.

Except for this weakness, CMOS is ideally suited for a timer: there is no input current and thus no need for Darlington stages.

Figure 11-9 shows a design using a more modern 5-Volt (0.5um) process. The comparators are conventional (as discussed in chapter 9), with the dimensions of the devices chosen so that the threshold and trigger inputs can move rail to rail and their matching is adequate for precision operation (3ppm/^{o}C).

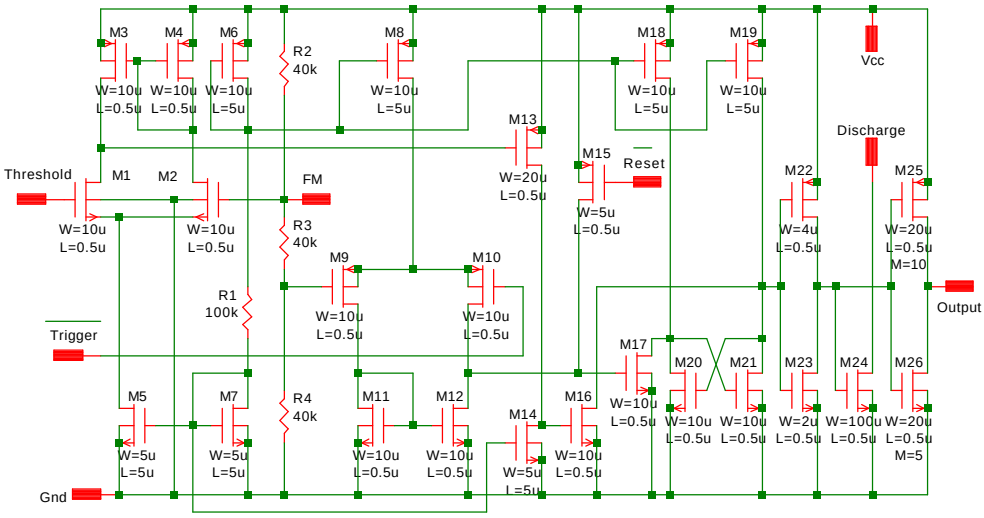


Fig. 11-9: A 5-Volt CMOS version of the 555 timer.

Ordinarily a flip-flop consists of two cross-coupled gates. In this case two cross-connected transistors fed by current sources result in smaller temperature and voltage drifts, because the flip-flop switch levels track the operating currents of the comparators.

The operating currents are set by R1, which limits the operating voltage range in which high precision is obtained to 3 to 5 Volts. Replacing R1 with a current source extends this range down to 1 Volt.

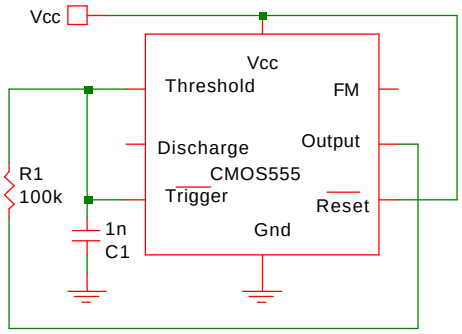


Fig. 11-10: 50% duty cycle oscillator.

A CMOS output stage swings rail to rail (unlike a bipolar output which at the very least has a minimum drop of some 150mV, if not an entire VBE). Thus the timing resistor can be connected to the output, resulting in a square-wave with a precise 50% duty-cycle. On the other hand, CMOS devices are inferior to bipolar ones when it comes to current handling. Even with a gate-width of 200um for the P-channel devices and 100um for the N-channel transistor, the circuit only delivers 10mA and the voltage drop is .25V (which badly affects the duty-cycle). It would be better to have separate outputs for the timing resistor and the load.

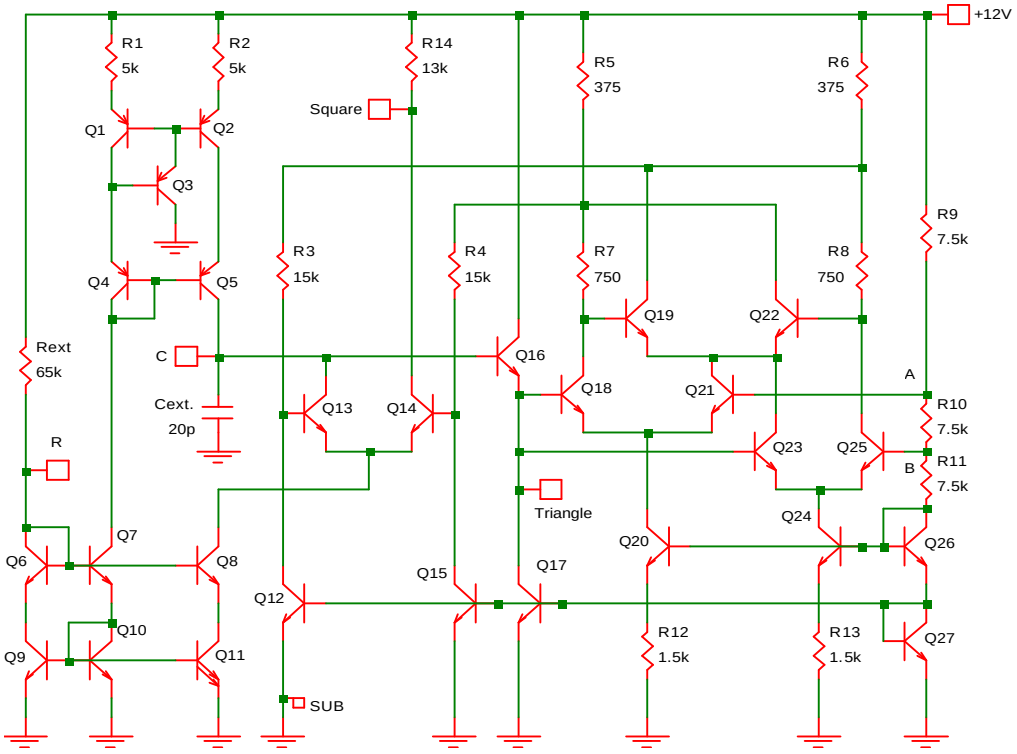


Fig. 11-11: A high-frequency triangle-wave generator.

Figure 11-11 shows an oscillator which produces a precise triangle waveform even at relatively high frequencies. All transistors which determine speed are NPN and do not saturate.

First the low-frequency part, the current sources used to charge and discharge the external capacitor. The primary current is produced by Rext; being connected between the positive supply and two VBEs the current through the resistor is not only dependent on the supply voltage but also has a temperature coefficient. Both of these effects are eliminated by using an internal resistor chain (R9, R10 and R11) connected the same way.

The primary current is mirrored by Q6, Q7, Q9 and Q10 and then mirrored again (Q1 through Q5) to form the charge current. A second current of twice the magnitude is derived from the first current mirror by Q8 and Q11. This latter current is used to discharge the capacitor and is turned on and off by the differential pair Q13/Q14.

The internal resistor chain is used to bias the rest of the circuitry and provide the reference voltages for two comparators; the voltage across the three identical resistors is (Vcc - 2VBE).

Comparator 1 consists of a single differential pair (Q18,Q21) as does comparator 2 (Q23, Q25). They provide the operating current for the flip-flop (Q19, Q22) with two of their collectors while the other collectors switch the flip-flop's bases. The ratio of the resistors in this arrangement is key: the operating currents for the comparators is set by R12 and R13, which have one VBE across them. The two currents end up flowing together through either R5 or R6, depending on the state of the flip-flop. R5 and R6 are one quarter the value of R12 and R13, so the voltage drop across them is one-half VBE. In other words, the collector voltages of the flip-flop drop 1/2 VBE below the base potential, which is safely above the saturation voltage.

Now the question is: how do we get this small voltage fluctuation, located just below Vcc, to work on the bases of the differential pair Q13/Q14, which must operate in the voltage region below the low point of the wave-form? We cannot use lateral PNP transistors, they are far too slow.

We do this by coupling the switching signal to the differential pair through two resistors (R3, R4) and running a known DC current through them. Q12 and Q15 are current mirrors, slaved to the bias chain (Q27). Their current thus increases as the supply is increased, and so do the voltage drops across R3 and R4. Thus the average potential at the bases of the differential pair stays at a fairly constant 1/3 Vcc, over an operating range from 9 to 15 Volts.

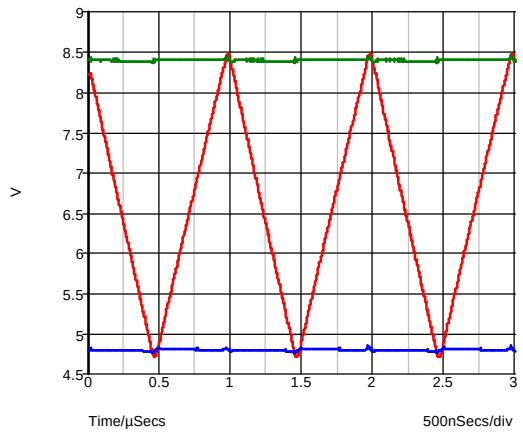


Fig. 11-12: Wave-form at 1MHz.

The triangle wave-form across the capacitor is buffered by the emitter follower Q16 for use by both the comparators and an external load. At the unused collector of the differential switching pair a square-wave can be obtained.

This is not an oscillator of ultimate precision, but it delivers a good-quality wave-form up to at least 1MHz. The temperature coefficient is 190ppm/^{o}C and the change in frequency from 9 to 15V supply is 1.7%. As always, these results are based on one particular process; it is a good idea to re-simulate the design for the process you are using.

A triangle-wave can be looked at as a sine-wave with distortion. This is not so far-fetched, because the distortion is only about 12%. If we round off the peaks, we end up with a fairly respectable sine-wave with relatively little effort.

Figure 11-13 shows a companion circuit to figure 11-11. It is inserted between points A and B, replacing R10. The triangle wave, entering through R1, encounters attenuators at six different voltage levels. Follow a positive-going signal: at first there is no attenuation; at about 0.6 Volts R11 kicks in, held at that level by Q6; at the next higher level R10 appears in parallel to R11 and at the final level an even smaller resistor, R2, reduces the signal even further.

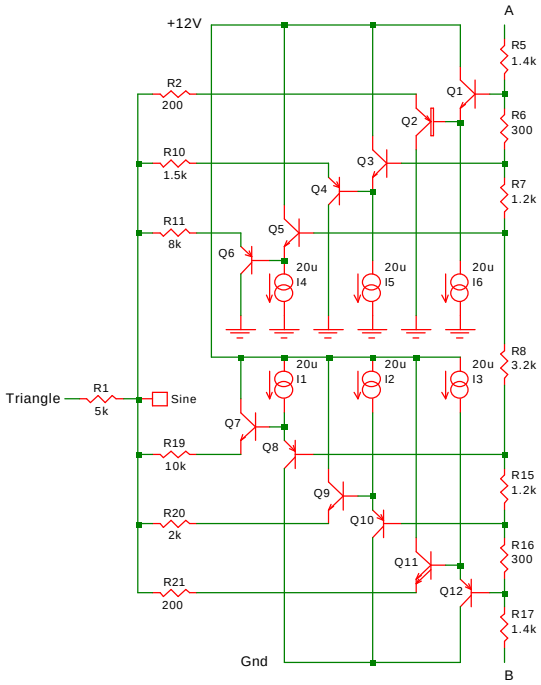


Fig. 11-13: Shaping circuit, transforming a triangle-wave into a sine-wave.

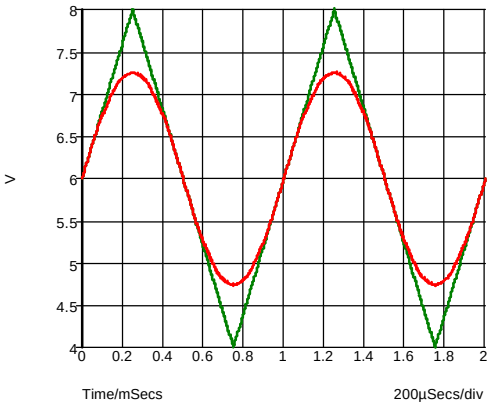


Fig. 11-14: Triangle and sine-wave.

Using only three clipping levels in each direction, the resulting sine-wave has a distortion of only 1%. Using more levels reduces the distortion, but is likely to require trimming.

Oscillators (and timers) can often be very simple for non-critical applications. Suppose you wanted to create a brief pulse once a second, for example to flash an LED. The frequency or pulse-width need not be precise, i.e. the design does not require two sophisticated comparators; a simple **Schmitt Trigger** will do.

Figure 11-15 shows a bipolar design for such an LED flasher, intended for use at either 5V or 3.3V. Q3 and Q4 form a simple comparator, with Q5 as an active load. Rext charges Cext and when the voltage at the base of Q3 exceeds that at the base of Q4, Q6, Q2 and Q7 turn on. Through the voltage divider R6/R7/R8 Q7 now abruptly lowers the potential at the base of Q4, while Q2 discharges Cext through R1. When the voltage across the capacitor drops to the new level at the base of Q4, Q6 turns off and the cycle starts anew.

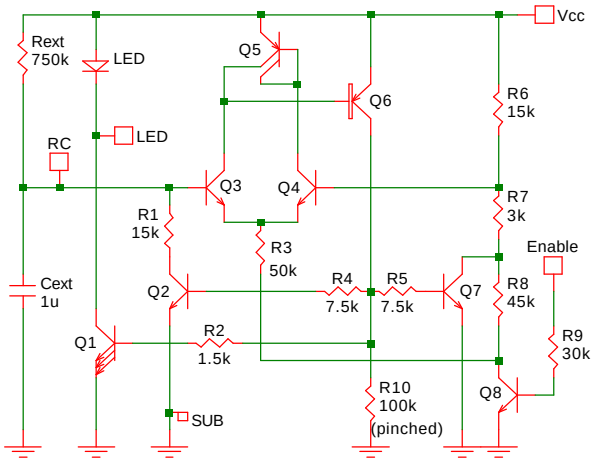


Fig. 11-15: Circuit generating brief current pulses.

The frequency is set by Rext and Cext (here 1 Hz) and the pulse-width by R1 and Cext (20msec). The frequency is accurate to within 2% from 3 to 5.5 Volts (not counting the variation of the external components), but the pulse-width reflects the variation of R1, a diffused resistor.

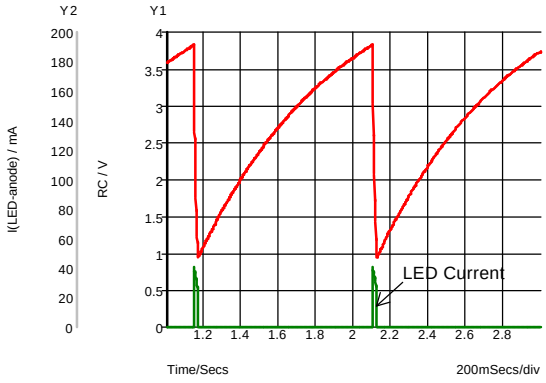


Fig. 11-16: Wave-forms of the pulse generator.

LEDs have a rather large forward voltage drop (about 2 Volts), so a supply voltage of at least 2.5 Volts is required. The current through the LED (40mA) is primarily determined by the size of Q1; it operates in the high-current region, where the gain has already decreased but the spread of hFE becomes narrow. Average current consumption is 1mA.

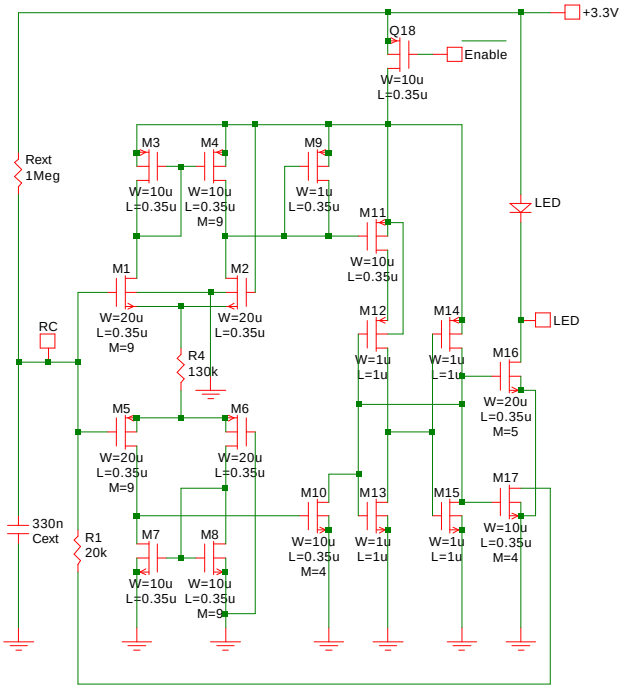


Fig. 11-17: CMOS design for a pulse generator.

It is interesting to consider a CMOS design for the same function. We can avoid using a resistive divider (and thus save current) by using two comparators and making two of the transistors in each comparator nine times the size of the others. This results in an offset of some 200mV. The reference potentials for the comparators are the supply lines. As the voltage across the external capacitor rises to 200mV below the positive supply, the upper comparator (M1 to M4) sets the flip-flop (M12 to M15); as it falls to 200mV above ground the lower comparator resets it. Cext is charged by Rext (1 second) and discharged by R1 (20msec). The two times are surprisingly accurate, exhibiting a 3% change from 3 to 3.6 Volts and 0 to 100°C.

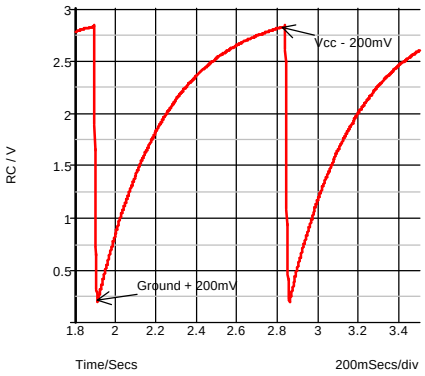


Fig. 11-18: Waveform at RC of figure 11-17.

The output current, on the other hand, shows the weakness of CMOS: it varies ± 21% with a supply voltage change of ± 10%. For this reason an even large output transistor (M16) and a resistor in series with the LED may have to be used.

With a 20msec pulse of 37mA every second, the entire circuit consumes just 650uA average.

A second example, a timer this time, shows just how low a power dissipation can be achieved if a resistor divider is avoided: the circuit in figure 11-19 draws 1uA at 1.8 Volts.

The entire circuit (save the output inverter) is powered from the Start pulse, rising from ground to Vdd. The logic input must stay high longer than the set timing. Alternately, if the Start terminal is simply connected to Vdd, the circuit becomes a start-up timer.

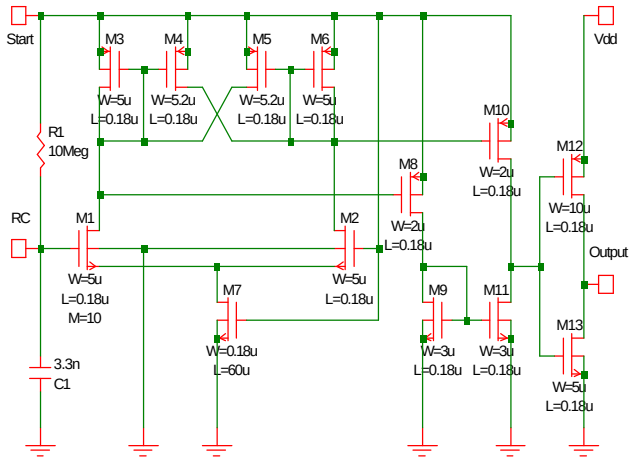


Fig. 11-19: A 1.8-Volt timer which consumes just 1uA.

M1 to M6 form a comparator. By making M1 ten times as wide as M2, an 80mV offset is created. The gate of M2 is connected at Vdd, thus the comparator switches at 80mV below Vdd. The switching action is enhanced by employing a small amount of positive feedback; M4 and M5 are slightly wider than M3 and M6 and deliver their drain currents to the opposite sides (see also figure 9-5)

The operating current for the comparator is provided by M7, a long and thin transistor. The advantage of such a device is size: it produces about 0.6uA using a relatively small area; a resistor doing the same job (1.2MOhms) would be painfully large.

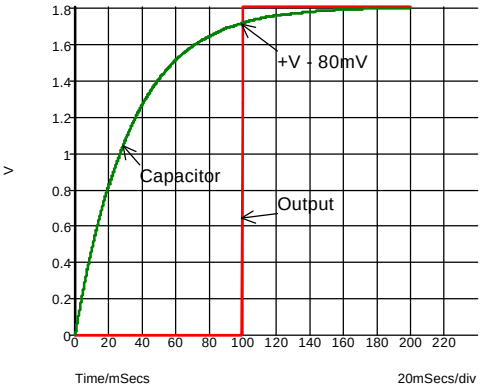


Fig. 11-20: Switching threshold of the 1.8V timer.

On the other hand an MOS transistor with its gate connected to the supply is hardly a constant current source. With a 25% change in supply (1.6 to 2 Volts) the current changes 70% (0.5 to 0.85uA). Here we can afford to live with this shortcoming.

The two outputs of the comparator are level-shifted by M8 and M10 and the active load M9/M11 to fit the input of an inverter.

A circuit operating at such a low current is of course quite slow; a significant timing error occurs below about 50usec. At lower speeds you can expect an accuracy of better than ± 5% from 1.6 to 2 Volts and 0 to 100^{o}C.

Simulation of Oscillators

You have just drawn up a great idea for an oscillator and start the simulation. Nothing happens, you get nothing but DC levels.

This situation is all too common. The simulator is trained to first find an operating point, i.e. set the DC voltages and currents so all the device equations are satisfied (in simulator-speak: find convergence). Then the transient analysis starts and the computer finds that all the voltages and currents remain unchanged over time.

In real life the circuit may start at exactly the same point, but no voltage or current stays unchanged, there is *noise*. No matter how tiny these fluctuations are, they move the circuit to a slightly different operating point and it becomes apparent that movement in one direction is the path to be followed. Thus, gradually, the oscillation builds momentum.

Without any noise (or some sort of transient disturbance) no circuit would oscillate; it would just sit there, precariously balanced.

It is of great advantage to have a simulator which allows real-time noise (i.e. all currents and resistors actually produce the appropriate amount of noise not only for a (small-signal) ac simulation, but during a transient analysis as well). With this feature a properly designed oscillator will always start. If you are using a simulator which has no real-time noise, you may have to coax the oscillator by jarring it with a pulse, e.g. step the supply voltage abruptly to a higher level.

But there is a second potential problem: *it may take a long time for the oscillation to build up.* For the circuits discussed so far in this chapter this is no great worry, but for the type of circuits we are about to encounter this can be very frustrating.

Take a crystal oscillator, for example. A high-quality crystal can take up to a second to start oscillating. At 10MHz, that amounts to 10 million cycles the simulator has to go through, in very small time steps to catch any movement. If you are not aware of this nuisance, you may sit there watching flat lines and come to the conclusion that your oscillator does not work.

LC Oscillators

Oscillators using inductors are rarely used in integrated circuits, except at frequencies above GHz. But there are occasions at lower frequencies when only an inductor can give you the performance required. One such example is shown in figure 11-21.

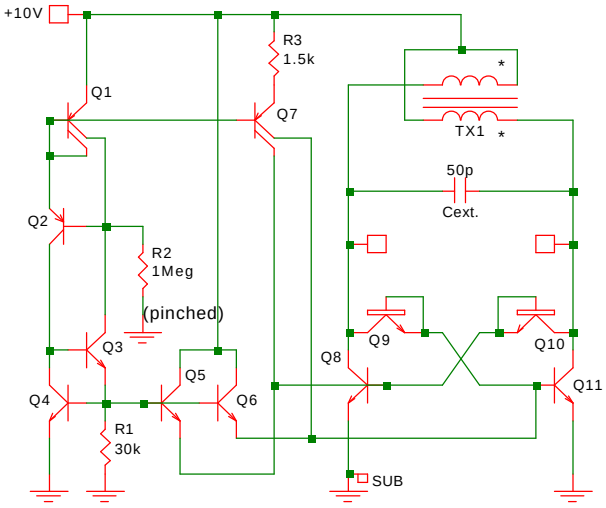


Fig. 11-21: Sine-wave oscillator with large amplitude.

The object of this design is the creation of a 10MHz sine-wave with a large amplitude. Q8 and Q11 are the oscillating transistors, cross-coupled by the collector-base capacitances of Q9 and Q10 (about 8pF each).

A small current (about 12uA) is injected into the bases of Q8 and Q11 to bring them into a current level at which there is sufficient gain. TX1 is in reality a center-tapped inductor, shown as a transformer with two windings for the simulation.

The collectors of the oscillating transistors start at the supply voltage, 10 Volts. After a few hundred cycles of gradually increasing amplitude the waveform at each collector is limited by the emitters of Q5 and Q6 at the negative end; at this point the peak-to-peak amplitude has reached 21 Volts, more than twice the supply voltage. The action of the center-tapped inductor is that of a see-saw: one end dips to ground (or slightly below) while the other peaks at a little above 20 Volts.

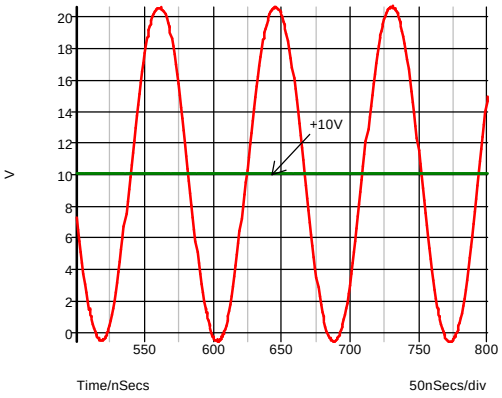


Fig. 11-22: The voltage swing of the oscillator extends to twice the supply voltage.

Though the supply voltage is 10 Volts, the four transistors Q8 to Q11 must have a voltage capability of 21 Volts.

To minimize the output capacitance, each oscillating transistor can share a collector region with its cross-coupling capacitor.

Crystal Oscillators

Let's start with the circuit commonly used in CMOS: the crystal is connected between the input and the output of an inverter. Since an inverter is ill-equipped to remain in a state between low and high, R1 is employed to force it into the linear region, at least initially. C1 and C2 are a mystery to most designers; they are there because the crystal manufacturer specifies them.

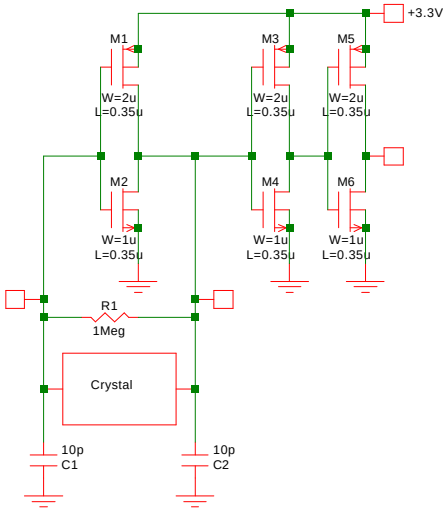


Fig. 11-23: CMOS crystal oscillator.

The whole arrangement is a bit curious. An oscillator needs to have positive feedback, yet the phase-shift between the input and the output of an inverter is 180 degrees - negative feedback. To understand this we need to look at the crystal itself.

A crystal is simply a sliver of quartz that vibrates. Quartz is piezoelectric, i.e. a voltage applied between two surfaces makes it flex and flexing it creates a voltage between its surfaces.

The vibrating mass of the crystal can be represented as a series-resonant LC circuit (C1, L1) with a series resistance R1. The Q (originally the quality factor) of such an LC circuit is given by:

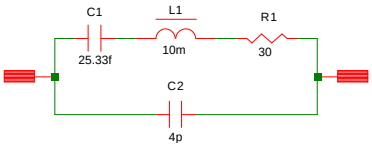


Fig. 11-24: Model for a crystal.

$$Q = \frac{2 * p * f * L_1}{R_1}$$

and the resonant frequency by:

$$f = \frac{1}{2*p*\sqrt{L_1 * C_1}}$$

The values in figure 11-24 were chosen to give a series-resonant frequency of exactly 10MHz and a Q of 20,000. The Qs of crystals range from 10,000 to 2 million, i.e. far higher than those of LC circuits. Ceramic resonators, which are otherwise almost identical to crystals, have considerably lower Qs.

C2 is the stray capacitance created by the contacts to the crystal and the wires and pins of the package.

If we open the feedback loop in the circuit of figure 11-23 (as shown in figure 6-14) we can see what is happening. There are in fact two resonances, about 0.2% apart. The lower one is the series resonance, the upper one parallel resonance. At these two frequencies the phase shifts abruptly between 180 and zero degrees.

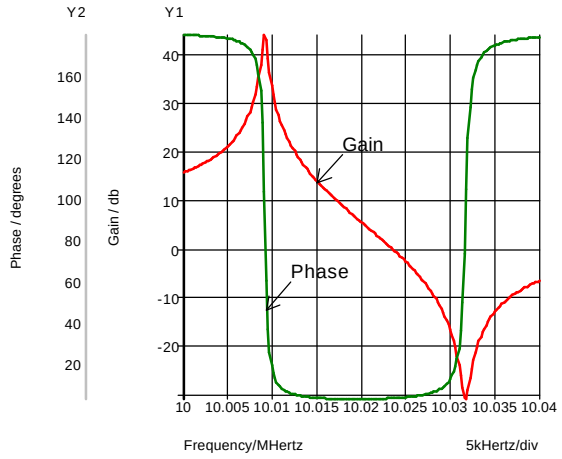


Figure 11-25: Series and parallel resonance of a crystal.

The parallel resonance is created by C1, C2 and the combination of external capacitances. In an oscillator properly designed for series resonance it is of no concern.

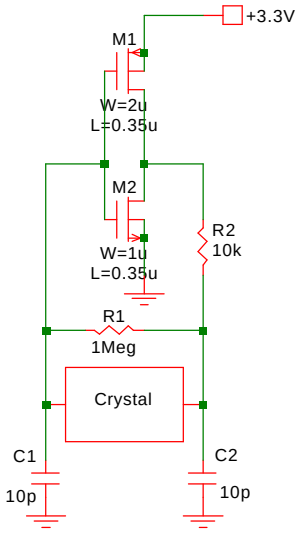


Fig. 11-26: Improved CMOS crystal oscillator.

But even the series resonance is influenced by the additional capacitances. As you notice from the plot it is not exactly 10MHz. There are two reasons: 1. At resonance the impedance of a series resonant LC circuit becomes very low, limited only by R1. Thus it works best if the input impedance of the inverter is low. In figure 11-23 all we have for input

impedance is a 10pF capacitance and the gate capacitances. Thus C1 sees itself in series with this capacitance and the effective capacitance is slightly smaller. 2. At resonance the phase moves from 180 degrees to zero, but it doesn't actually reach zero (the condition required for oscillation) until about 10.015MHz.

If we shift the phase in the feedback loop with an additional resistor (working against C2) a zero degree phase-shift is reached at the series-resonant frequency. With R2 the frequency is more accurate and the chances of the crystal operating at some unwanted frequency (including harmonics) are diminished. But be aware that R2 decreases the loop gain; make sure it safely exceeds unity.

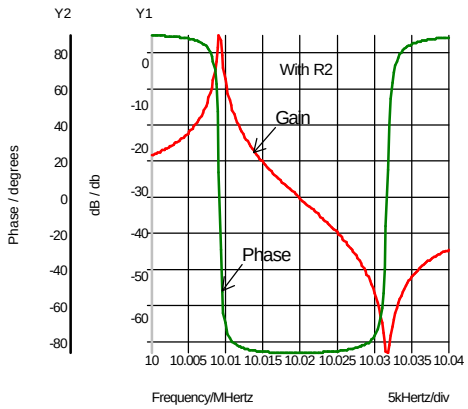


Figure 11-27: The insertion of R2 in figure 11-26 brings the phase shift to zero degrees at the series resonance.

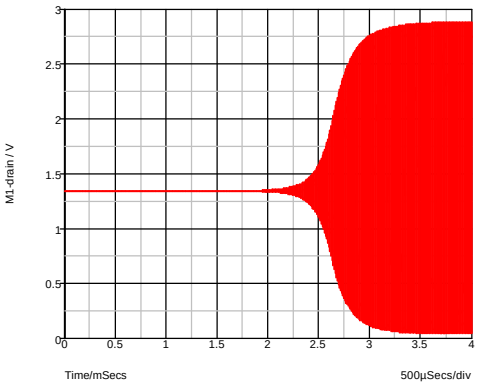


Figure 11-28: Start-up of a crystal oscillator.

Simulating a crystal oscillator can be a frustrating task. The higher the Q, the longer it takes for the oscillation to build up. In the case of figure 11-26 it takes 3msec for the oscillation to reach full amplitude. If you want to measure the frequency accurately, you need to do this in very fine steps (say 1nsec), i.e. you have to wait for 30,000 cycles before you can see the actual wave-form.

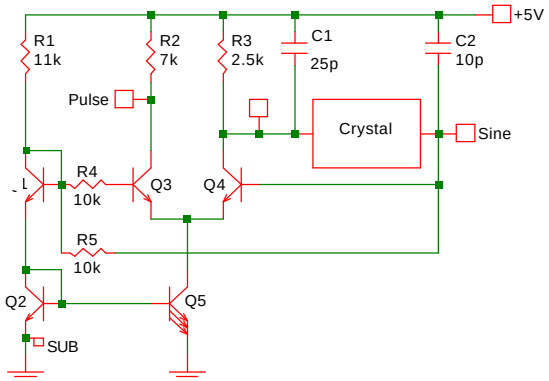


Fig. 11-29: Alternate crystal oscillator with bipolar devices.

Figure 11-29 shows a different approach, for a bipolar process. Gain and a 180 degree

phase-shift is obtained through Q4, which needs to run at a fairly substantial current (about 1mA for the differential pair). The base of Q4 is biased at 2VBE, which gives sufficient voltage swing without saturating the transistor.

Q3, on the other hand, saturates. By making its collector resistor larger than that of Q4, a pulse (square-wave) output is obtained, swinging between 0.5 and 5 Volts.

12 The Phase-Locked Loop

The idea of the phase-locked loop surfaced as early as 1932, but remained an esoteric and expensive concept until the arrival of the integrated circuit. The PLL has the unique ability to capture a signal without requiring precision components, a welcome feature in a world of large variations.

The phase-locked loop is primarily an analog concept and only when we treat it as such can we fathom its powerful capabilities. And here is where simulation fails us. Anyone who has ever sat down at a bench and observed a phase-locked loop grabbing a minute signal seemingly buried in noise will agree that a simulation cannot give you the same sensation. In a real circuit, a phase-locked loop almost seems to be alive, capturing and hanging on to a signal as the frequency is changed; a simulation of the same circuit is a cold experience, giving you no intuition (and taking up an enormous amount of time).

To get a feel for the operation of a phase-locked loop you need to understand the key component, the phase detector. So, let's begin with this.

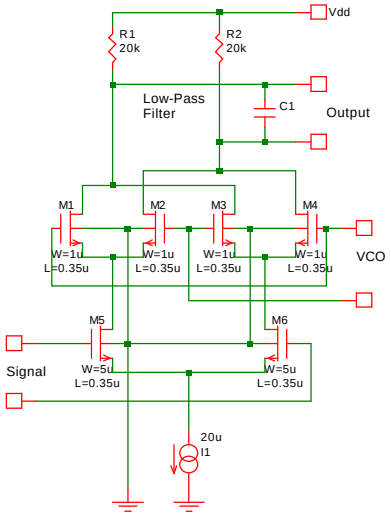


Fig. 12-1: The phase detector.

The circuit in figure 12-1 is known as the **four-quadrant multiplier**, one of several schemes used as phase-detectors.

There is a straight-forward differential pair, M5 and M6, which amplifies a signal arriving from outside the IC. But instead of the drains being connected to load resistors or an active load, their currents pass through four other transistors, which are turned on and off by a local **voltage-controlled oscillator** (VCO).

The signal inputs are at a certain DC bias level, say 1V or 1.5V, high enough to exceed the threshold voltages.

Now imagine a square-wave at the

VCO input with a frequency exactly the same as that of the input signal. This could be a rail-to-rail 2-phase square-wave or, preferably, a wave-form moving ± 200mV differentially, centered at about 2V DC. At first (figure 12-2) the VCO wave-form is *in-phase* with the input signal, i.e. the two cross zero at the same time.

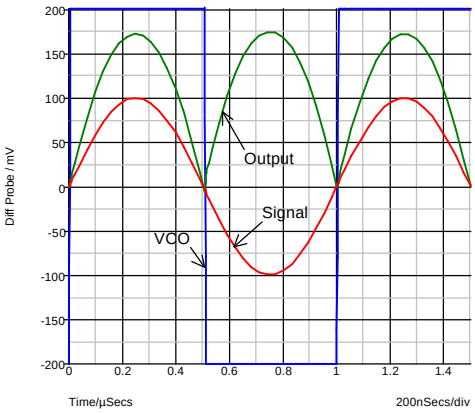


Fig. 12-2: With the VCO signal in-phase, the output is a rectified sine-wave with a positive average.

During the first phase of the VCO signal, the drain of M5 is connected to R1 (through M1) and the drain of M6 to R2 (through M4). During the second phase this connection reverses: The drain current of M5 flows through M2 and R2 and that of M6 through M3 and R1. Thus, ignoring C1 for now, the output across the two load resistors is a rectified sine-wave with a positive average value.

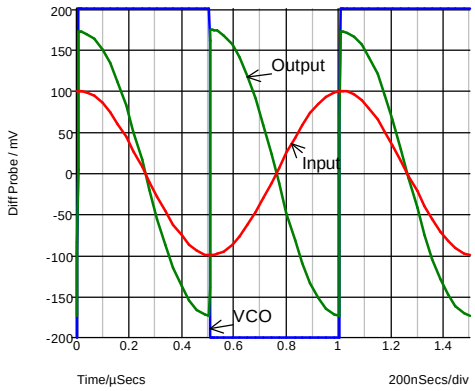


Fig. 12-3: With the phase of the VCO signal shifted by 90 degrees, the average of the output voltage is zero.

Now, let's keep the frequencies the same but shift the phase of the VCO signal by 90 degrees (figure 12-3). The signal is now chopped at the moment it reaches its peak amplitude and the output shows equal positive and negative excursions. Thus the average differential output voltage is zero.

If we shift the VCO wave-form a further 90 degrees (still keeping its frequency constant), the VCO wave-form chops the signal at the zero-crossings again, but now the output is inverted. Averaging it with C1 results in a negative voltage.

Thus, with C1 back in the circuit, we have a DC (or low-frequency) signal at the output of the phase detector which is a measure of the relative phases of the two frequencies. If we use this "error" signal to adjust the frequency of the voltage-controlled oscillator, we have the phase-locked loop (figure 12-5).

Here is what will happen in slow-motion: Let's say the VCO is running at 1MHz and the input signal is some distance away in frequency, e.g. 800kHz. Since the two frequencies are not synchronized, there is no phase relationship yet. At this point the phase detector is merely a mixer, producing several new frequencies, such as the *difference* between the two frequencies and various combinations of harmonics. The

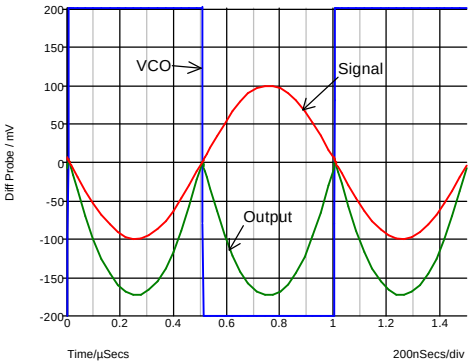


Fig. 12-4: At 180 degree phase shift the average of the output is negative.

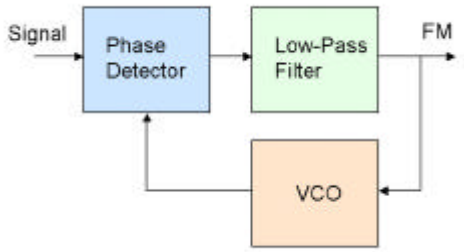


Fig. 12-5: Block diagram of a simple phase-locked loop.

AC, but the VCO starts to jitter around its free-running frequency. Move the input signal just a little higher in frequency and suddenly the jitter disappears and the VCO jumps into step with the input signal. Now we see a DC level at the output of the filter.

As you continue to move the input signal higher in frequency, the VCO will continue to track it, until the loop finally runs out of control voltage. This is illustrated in figure 12-6 where signal frequency is swept from low to high over a 5msec period.

one of interest is the difference, 200kHz; it is still too high to pass through the filter.

As we move the input signal gradually higher in frequency, there comes a point where the difference frequency is low enough so that some of the signal passes through the filter and starts influencing the VCO. The signal is not rectified yet, it is still

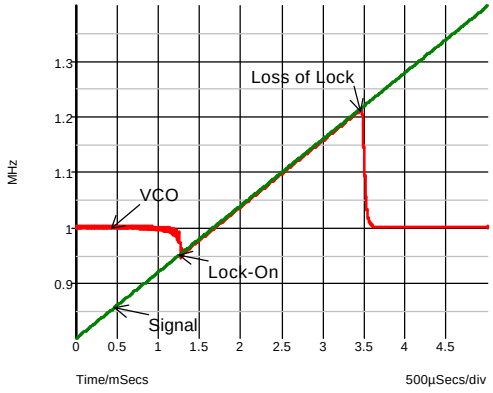


Fig. 12-6: Locking behavior of a phase-locked loop.

The exact same behavior is seen as you approach the VCO frequency from the high end. The capture range (the maximum difference between the two frequencies to *achieve* lock-on) is always narrower than the lock range (how far you can drag the VCO and still *keep* lock). It makes no difference which frequency is moved and which is fixed.

Both the capture range and the lock range are influenced by loop gain and signal level. If you increase the gain of the loop or the input signal level, both ranges become wider.

(Strictly speaking the name phase-locked loop is a misnomer. As you move through the lock range, the two frequencies are locked but their phase relationship *has* to change to produce the error signal, i.e. the phase is not locked. Frequency-locked loop would be a better name).

The phase-locked loop has three important features, especially for integrated circuits:

1. Apart from the loop gain, the capture range is determined by a single low pass filter. For example, if the signal you are looking for is at 50Mhz and has a narrow band-width (say 5kHz), you dimension the low-pass filter so it rolls off at about 5kHz. This makes the phase-locked loop look like a very sharp band-pass filter. A single-pole low-pass filter rolls off at 20dB per decade, so at 49.9MHz and 50.1MHz the interfering signal is attenuated at the low-pass filter by 26db. Using an active filter, it would take many poles and a large number of precision components to achieve the same selectivity.

The phase-locked loop depicted in figure 12-5 is a second order PLL (i.e. it has two poles, one by the VCO itself, the other by the low-pass filter). This configuration is unconditionally stable. Adding another pole makes stability (i.e. the absence of unwanted oscillation) more difficult to achieve, but it doubles the sharpness of the filter action.

2. The VCO need not be highly accurate. As long as the free-running frequency is within the capture range of the signal, the loop will find the exact frequency.

This advantage, however, is made a bit difficult if your capture range is very narrow. In the example above the free-running frequency would have to be within 5kHz of the signal, i.e. 0.1%. Without using accurate components, such precision can only be achieved by tuning, for example by sweeping the VCO over a wider range, detect capture (see below) and then stop the sweep.

3. The error signal (i.e. the output of the low-pass filter) is a measure of frequency deviation. If the input signal is frequency modulated, this output *is* the demodulated signal.

There is even a simple solution if the modulation is AM, not FM (figure 12-7). In this approach the VCO has a second output (same frequency but shifted 90 degrees) and a second phase detector and low-pass filter are added.

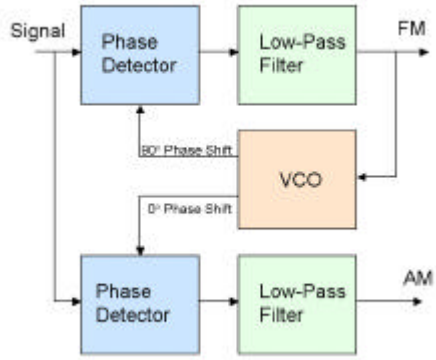


Fig. 12-7: Phase-locked loop with AM output.

In the middle of the lock range, the phase shift between the signal and the VCO is automatically 90 degrees, so that the control voltage for the VCO is zero. This means that the signal frequency is chopped at the amplitude peaks. A second phase detector operating at zero degrees phase shift will, therefore, chop the same signal at the zero crossing and the result is a voltage proportional to amplitude. This output then delivers not only the demodulated AM but also indicates that the loop is locked.

How do you design a voltage-controlled oscillator? We have seen some examples in chapter 11, but for most applications VCOs for phase-locked loops are specialized for high-frequency operation. We have two examples here.

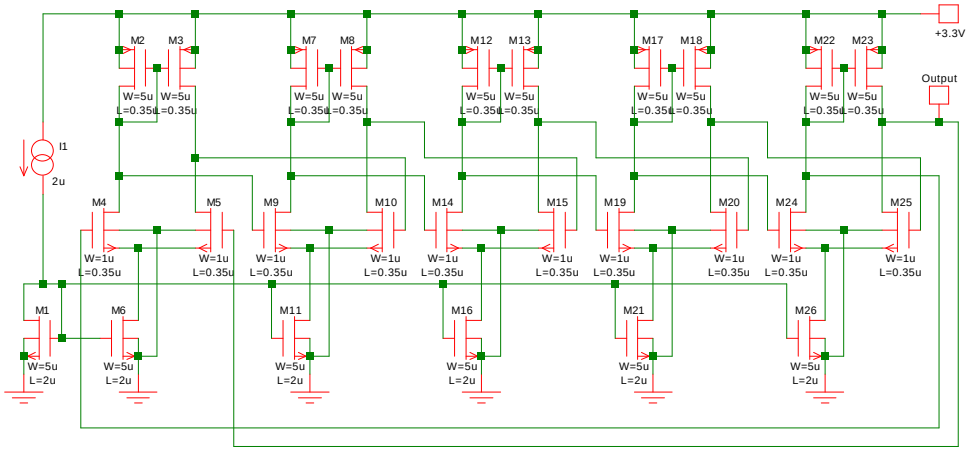


Fig. 12-8: Current-controlled oscillator for operation between 6MHz and 300MHz.

The first example is a current-controlled oscillator, though it is a simple matter to convert the current into a voltage.

I1 sets the operating currents for five simple differential amplifiers with active loads. The outputs of each amplifier are connected to the inputs of the next stage and the outputs of the last stage back to the inputs of the first. It is a **ring oscillator**, relying on the delay caused in each stage. This delay is dependent on the operating current, increasing as the operating current is decreased. With I1=10uA the delay in each stage amounts to 1.6nsec and the frequency is 300Mhz. With I1=0.1uA the delay increases to 83nsec and the frequency decreases to 6MHz. A remarkably large range.

But the delay in each stage is caused by a number of effects, each with its own temperature coefficient. The net result is a variation in temperature coefficient from +800ppm/^{o}C at 0.1uA to +200ppm/^{o}C at 10uA. This temperature coefficient can be partially compensated by introducing an opposite tempco into the voltage to current converter, most likely optimized at one operating current only.

Our second example (figure 12-9) uses a different approach. Similar to the 566 oscillator in figure 11-1, a capacitor (C1) is charged and

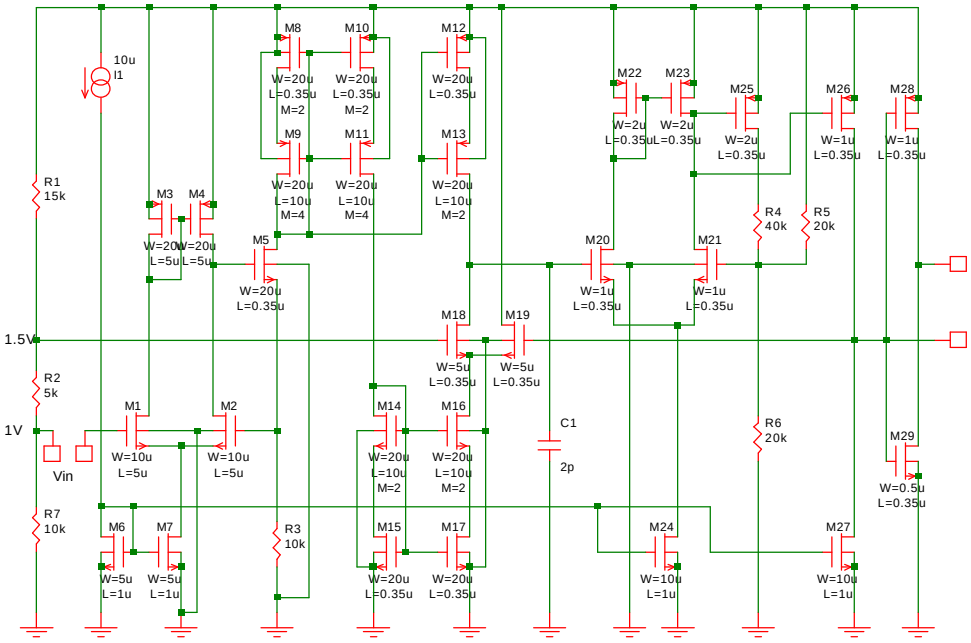


Fig. 12-9: Voltage-controlled oscillator with a Schmitt Trigger.

discharged by a current. But using two comparators and a flip-flop would be too slow for high-frequency operation. So we employ a **Schmitt Trigger**, which has fewer devices in sequence and thus reduced delay.

There are two thresholds again. The lower one is 1/2*Vdd and is set by the two equal resistors R5 and R6. When this lower threshold is reached, M21 and M25 turn on. This connects R4 in parallel to R5, which makes the upper threshold 2/3*Vdd. Notice that the resulting triangle waveform at the capacitor has an amplitude of only 1/6*Vdd peak-to-peak; we are trading accuracy for speed.

This is a somewhat improved version of a Schmitt Trigger; the most important factor determining accuracy is the "ON" resistance of M25. If it amounts to a substantial part of the value of R4, the effective resistance will be higher and will have a different temperature coefficient than R5 and R6. To make this "on" resistance small, we should increase the gate width of M25, but we can only do this at the expense of speed. The dimensions chosen for M25 are a compromise.

There is a separate stage (M26) to create a rail-to-rail swing and an inverter (M28, M29) to make both phases available to the phase detector.

The rail-to-rail output of the Schmitt Trigger is also used to switch the capacitor current between charge and discharge (M18, M19). Again, the dimensions chosen here are a compromise. For accuracy over a wide control current range we want them to be large; to get a fast response, they need to be small.

There is a voltage to current converter (M1-M7) and R3 is intended to be an external resistor. The control voltage is derived from Vdd through a resistor divider (R1, R2, R7) with a rest value of 1 Volt, so that the current tracks the two thresholds and the frequency is independent of supply voltage. With no signal at the input, the voltage to current converter produces 100uA. A large-value resistor can be inserted between the two input terminals and the base of M1 modulated with the error signal (thus changing the current by perhaps ±10uA or ±20uA. To adapt the phase detector shown in figure 12-1 to this VCO, an active load can be used, converting the differential error signal to a single-ended one and then bringing it to near ground potential with a current mirror.

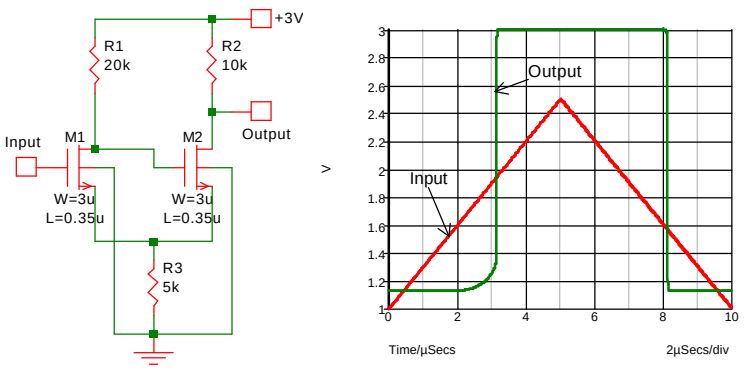
With C1=2pF the frequency of oscillation is 36MHz, with a temperature coefficient of -370ppm/^{o}C. At 60MHz (C1=1pF) the temperature coefficient increases to -680ppm/^{o}C because of the greater role played by delay. Below 20MHz the temperature coefficient is close to zero. There is a ± 0.3% change in frequency for a ± 10% change in supply voltage.

The Schmitt Trigger

Otto H. Schmitt was not a German electronics engineer as one would expect. He was born in St. Louis, Missouri in 1913 and studied biophysics and zoology. All through his life he built the electronic equipment he needed for his research himself and became an expert in electronics as well. In 1934 (at age 21) he was engaged in the study of the nervous system and came up with a bistable circuit which mimicked the behavior of a nerve (using vacuum tubes, of course). He never patented it.

The schematic below is a translation of his circuit into MOS devices. With the input low, the gate of M2 is biased high through R1, thus M2 is turned on. The ratio of R2 to R3 sets a bias point at the sources of the two transistors. When the input is moved above this point (plus the threshold voltage of M1), M1 turns on, M2 turns off and the bias voltage is set to a lower level by the ratio of R1 to R3.

Otto Schmitt died in 1998, after a long and productive tenure at the University of Minnesota. Although he is best-known for his "Schmitt Trigger", it represents only a minute fraction of his contributions to science and engineering.



13 Filters

We can go back as far as 100 years and find elaborate electronic filters, using inductors, capacitors and resistors. And the inductor in these combinations has always been the problem child, the largest, heaviest, most expensive and least reliable component. With the advent of integrated circuits its status moved from undesirable to virtually impossible.

There is an intriguing relationship between the inductor and the capacitor; they are direct opposites. As you charge an inductor, the voltage appears first, the current follows later; in a capacitor the current must flow before the voltage can build up. If we build a circuit which shifts the phase 180 degrees, a capacitor has all the appearances of an inductor. It is on this phenomenon that IC filters are based.

Active Filters: Low-Pass

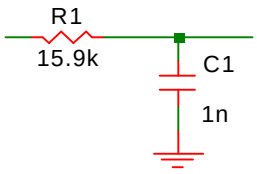


Fig. 13-1: Single-pole RC low-pass filter.

Consider a simple RC network. It has a cutoff frequency (the point at which the amplitude drops by 3dB) of:

$$f_{3dB} = \frac{1}{2pRC}$$

For the values shown in figure 13-1 this amounts to 10kHz. Below about 1kHz there is no attenuation. At 10kHz the signal at the output is down by 3dB and at 100kHz (10 times f_c) the attenuation amounts to 20dB. If you extend the straight portion of the curve (figure 13-2) upward it points precisely at 10kHz. Such a single-pole, passive RC

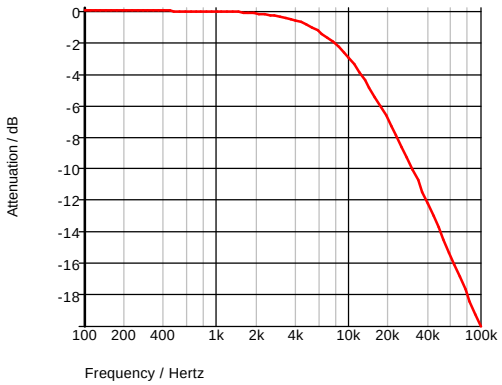


Fig. 13-2: Frequency response of a single-pole filter.

low-pass filter is said to have an attenuation of 20dB per decade or 6dB per octave (doubling of the frequency).

We don't need to be satisfied with just one RC network, we can connect several of them in series (i.e. cascade them). But we need to put a buffer in between the stages, otherwise the network following will load down the previous one too much.

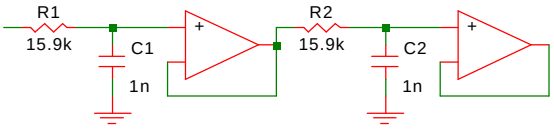


Fig. 13-3: A poor way to sharpen filter response.

But look at what we have done (figure 13-4). A second stage will roll off faster, but it also lowers the -3dB frequency. The more stages there are in series, the lower the cutoff frequency. With 16 stages it has moved down to a little over 2kHz.

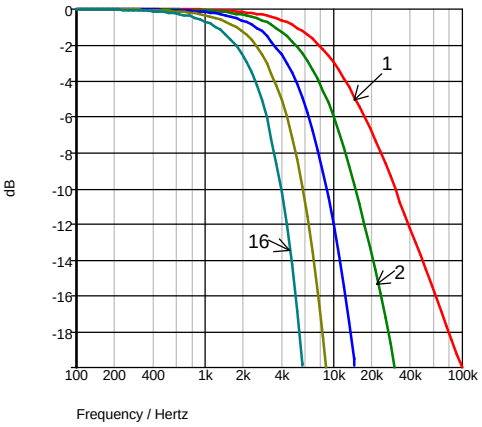


Fig. 13-4: Placing identical low-pass stages in series lowers the cutoff frequency.

We can do much better than this on two fronts. First, the frequency response can be shaped at will by choosing different resistor and capacitor values for each stage. There are schemes for this, worked out mathematically a long time ago, in the era of passive filters. They carry such names as Butterworth, Bessel and Chebyshev.

Second, we can take advantage of active components (such as op-amps) and create more compact filter stages. There are many such designs, with names such as Sallen & Key, Multiple Feedback, Fliege, Bach, KHN, or Tow-Thomas.

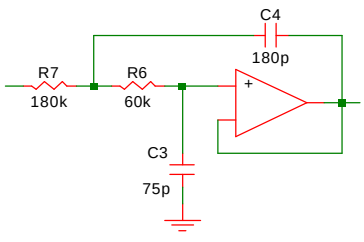


Fig. 13-5: Second order Sallen & Key Butterworth filter.

Figure 13-5 shows a second-order active filter, using a design developed by R.P. Sallen and E.L. Key. Only one op-amp is required for two poles.

The component values are chosen to give a Butterworth response; with different values we could change the frequency response to a Bessel or Chebyshev function.

As you can see from figure 13-6, the drop-off is now twice as steep as that of a

single RC network (i.e. 40dB per decade or 12dB per octave) and the -3dB point has remained at 10kHz.

Let's now take a look at three filters. The nominal designs are identical; they all have two cascaded second-order Sallen & Key stages. But each filter has different R and C values.

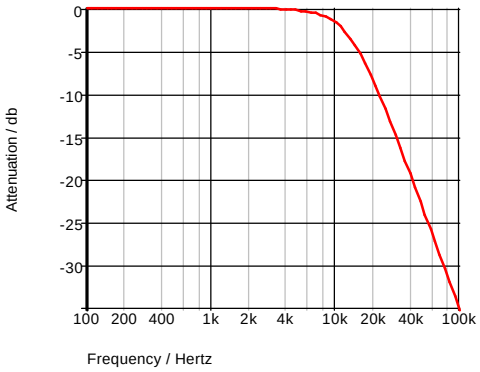


Fig. 13-6: Frequency response of the filter in figure 13-5.

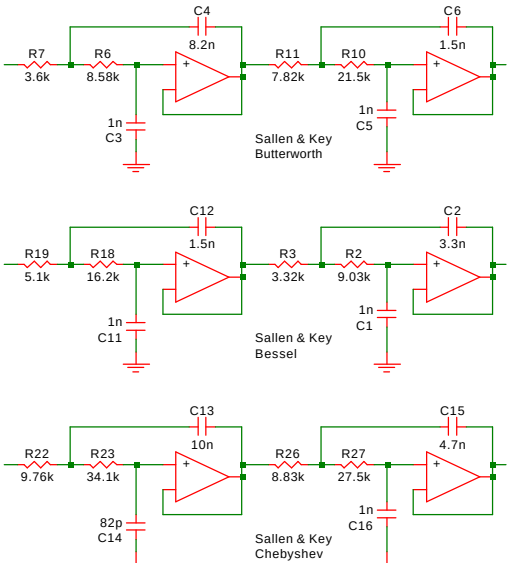


Fig. 13-7: Three fourth-order low-pass filters. The different component values result in different frequency responses.

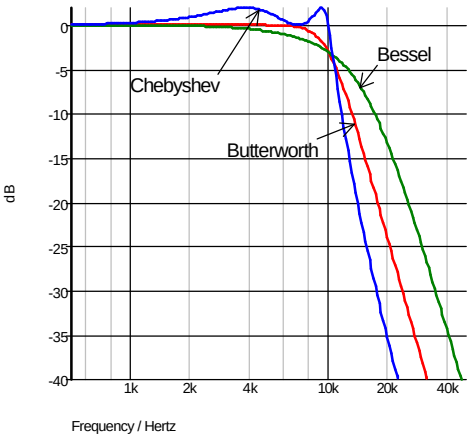


Fig. 13-8: Frequency responses of the three low-pass filters.

Judging by the frequency response alone, the Chebyshev filter has the sharpest response, though it produces some ripples in the pass-band (i.e. below 10kHz). This ripple can be reduced, at the expense of steepness above 10kHz. In even-order Chebyshev filters the ripples are above the line (0dB in this case); in odd-order ones they are below the line.

The Bessel filter gives a gentle roll-off with no overshoot in the pass-band, and the performance of the Butterworth filter is in between the other two.

But there is more to the performance of a filter than just the frequency response. Take the phase of the signal, for example. It never stays constant in any filter with the delays caused by the capacitors. But there is a difference between the three filter types. The Bessel filter has the smallest phase-shift, the Chebyshev the largest.

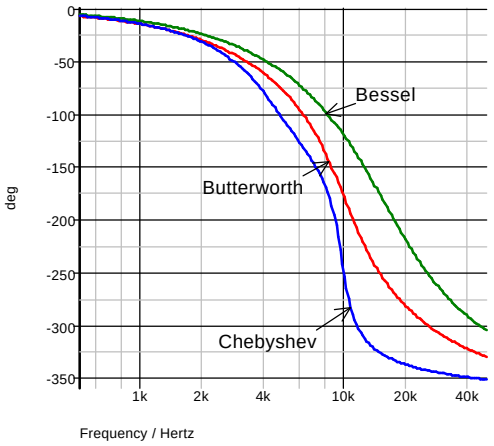


Fig. 13-9: Phase response of the three filters.

The phase response influences two more measures of filter quality. The first one is called **Group Delay**, shown in figure 13-10. Assume that you pass through the filter not just one frequency, but several. A delay in the filter causes the phase-relationships of the different frequencies to change and distortion results.

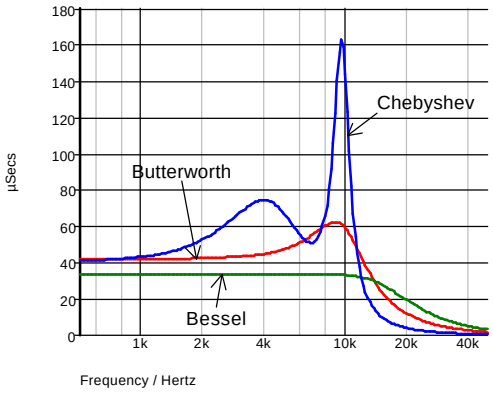


Fig. 13-10: Group delay of the three filters.

The Bessel filter is by far the best in this respect, having not only the shortest delay but also the most constant. The Chebyshev filter is by far the wildest.

Also, we can judge a filter by its pulse response. In figure 13-11 a 100usec pulse was applied to the input. We expect a rounding of the corners at the output but, considering that all three filters have the same cut-off frequency, the Bessel filter does the best job.

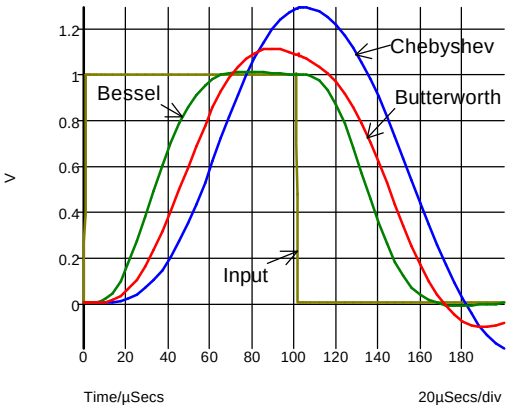


Fig. 13-11: Pulse response of the three filters.

How do we get the values for the resistors and capacitors? If you open up a text-book on filters, you will

see elaborate tables giving you coefficients for Butterworth, Bessel and Chebyshev functions. This is no longer necessary. There are a multitude of programs available on the web (many of them at no cost), which calculate these values for you. Search for "active filter software".

Bessel, Chebyshev and Butterworth

Friedrich Wilhelm **Bessel** (1784 to 1846) was a professor of astronomy at the University of Königsberg in Germany. By measuring the position of some 50,000 stars he greatly advanced the state of celestial mechanics and came up with the Bessel function, which was found to be also useful in filters.

Pafnuty **Chebyshev** (1821 to 1894) taught mathematics at the University of St. Petersburg. His major contribution was the theory of prime numbers but, similar to Bessel he left behind a function which later turned out to be applicable to filters.

Of Stephen **Butterworth** we know only that he worked at the British Admiralty for almost all his life. In 1930 he published a paper "On the Theory of Filters". He died in 1958.

Let's look at two more low-pass filters, using designs other than Sallen & Key. The two stages in figure 13-12 use voltage-controlled voltage-sources (VCVS), an approach differing from Sallen & Key only in that the op-amps have gain.

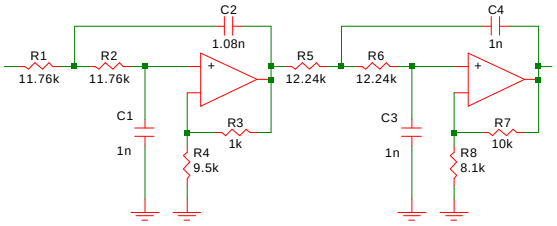


Fig. 13-12: 4th-order low-pass Butterworth filter in a voltage-controlled voltage-source design.

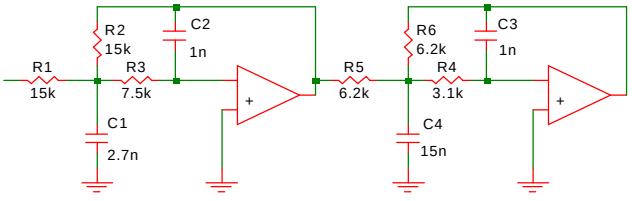


Fig. 13-13: A 4th-order Multiple Feedback approach.

The design approach for each stage of figure 13-13 is called Multiple Feedback.

All these different approaches render the same frequency and phase response, but they differ in sensitivity, i.e. how much component and op-amp parameter variations will influence filter performance. A temperature and Monte Carlo analysis reveals the merits.

High-Pass Filters

There is no mystery to converting a low-pass filter into a high-pass one: you simply exchange resistors and capacitors.

The drop-off now occurs toward the low-frequency end, but at the same rate as that of a low-pass filter, 80dB per decade for a fourth-order filter.

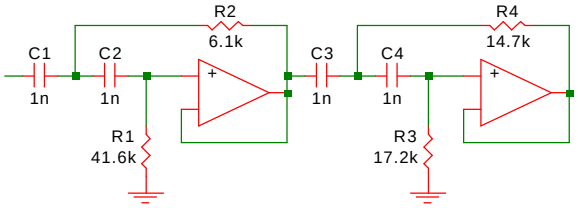


Fig. 13-14: High-pass Sallen & Key filter with Butterworth values.

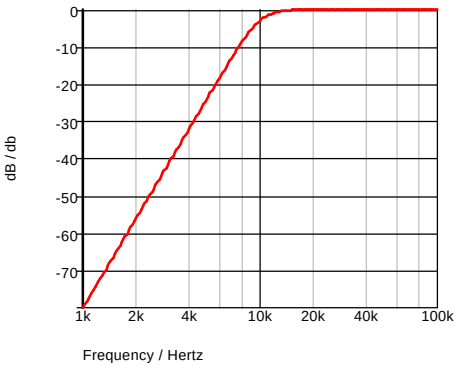


Fig. 13-15: Frequency response of the 4th-order high-pass filter of figure 13-14.

Note that in all of these drawings, abstract op-amps are used (inside the symbol is an ideal voltage-controlled voltage-source). In a practical design you have to consider the power supply. With a single supply, you may have to bias the input midway between ground and +V. In figure 13-14 this is accomplished at the low ends of R1 and R3.

Band-Pass Filters

Take the second-order low-pass filter of figure 13-5 and convert one RC network to high-pass. You now have a drop-off in amplitude at both high and low frequencies.

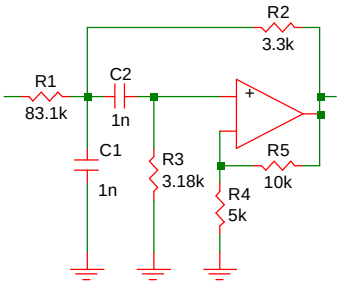


Fig. 13-16: Sallen & Key band-pass filter.

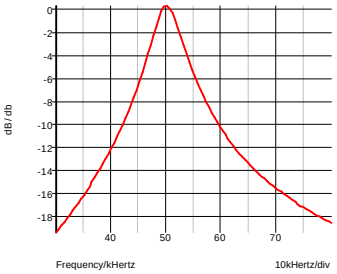


Fig. 13-17: Second-order band-pass filter response.

Although the arrangement is called a second-order band-pass filter, the drop-off rate is only first-order, 20dB per decade, since only one pole is active in each frequency segment. We can of course improve this by adding more stages, each stage contributing another 20dB per decade drop-off. And here there is a bewildering number of schemes available, with names such as Wien-Robinson, Deliyannis, Fliege, Twin-T, Mikhael-Bhattacharyya, Berka-Herpy and Akerberg-Mossberg. Your filter program will tell you which one to choose.

There is also an additional choice for the frequency response: compared to the Chebyshev filter the elliptic (or Cauer) has an even steeper initial drop-off, but the attenuation in the stop-band (i.e. outside the pass-band) is not flat.

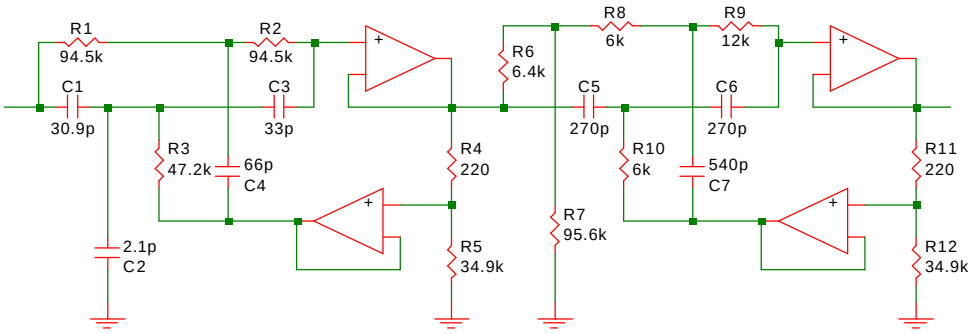


Fig. 13-18: A fourth-order, twin-T elliptic band-pass filter.

The filter of figure 13-18 has two **Twin-T** stages. The first stage is a second-order low-pass notch configuration, the second stage is called a second-order high-pass notch filter. The center frequency was chosen to be 50kHz, the bandwidth 2kHz. Just outside the bandwidth the attenuation reaches a maximum, but then settles down to a modest 15dB.

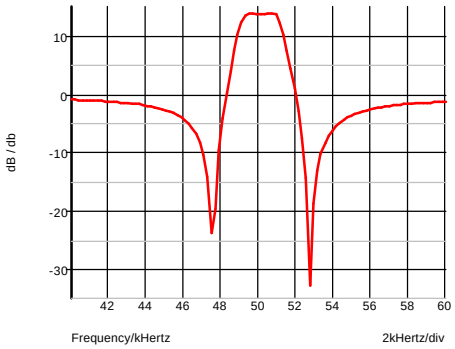


Fig. 13-19: Response of a fourth-order elliptic band-pass filter.

It must be clear to you by now that active filters are costly. Not only do they require precision components, but the values of most capacitors and some of the resistors are such that they cannot be integrated. A fourth-order low-pass or high-pass filter requires at least eight external

components and five pins. For a band-pass filter with only modest performance 14 external components and pins are needed.

Switched Capacitor Filters

If we charge a capacitor (C_R) by closing switch S1 for a brief period of time, then open S1 and close S2 for the same amount of time, the potential across the capacitor is first that of V1, then V2.

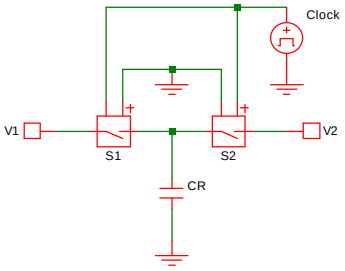


Fig. 13-20: Making a resistor out of a capacitor by switching at a rapid rate.

One of the handiest formulae to carry in your mind is:

$$Q = C*V = I*t$$

i.e. the charge in a capacitor (in Coulombs) is given by either the capacitance times the voltage or the current flowing into the capacitor for a certain period of time. In the case of figure 13-20, the current flowing between the two terminals over one period is

$$I = \frac{C_R*(V1-V2)}{t_{clock}} = C_R*(V1-V2)*f_{clock}$$

If we had a resistor between V1 and V2 instead of the switches and the capacitor, the current flowing through it would be:

$$I = \frac{(V1-V2)}{R}$$

Thus the equivalent resistance of the switched capacitor is:

$$R = \frac{1}{C_R*f_{clock}}$$

Let's look at some numbers. Suppose the switching frequency is 100kHz and C_R = 5pF:

$$R = \frac{1}{10^5*5*10^{-12}} = 2*10^6 = 2\,MegOhms$$

Thus, with a relatively small capacitor we can create the equivalent of a large-value resistor. If we were to implement such a device directly, the cost in area would be prohibitive.

But the area reduction is just the first benefit of switching; there is more: if we use this resistor in a filter, the absolute capacitance value disappears.

Shown here is a simple, one-pole low-pass filter. The cutoff frequency is given by:

$$f_{3dB} = \frac{1}{2*p*R*C}$$

Substituting the equivalent resistance of the switched capacitor we get:

$$f_{3dB} = \frac{f_{clock}*C_R}{2*p*C}$$

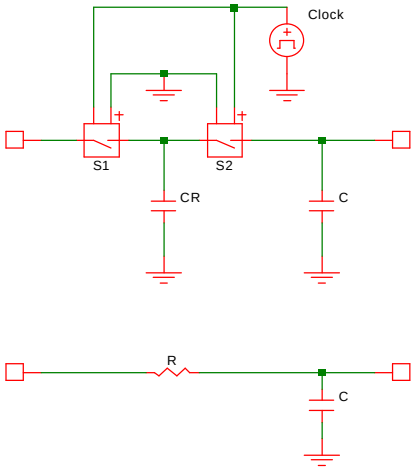


Fig.13-21: By using the equivalent resistance of a switched capacitor in a filter, only the capacitor *ratio* and the clock frequency are important.

If we make the two capacitors equal (any value) and switch at a rate of 100kHz we get a filter with a cutoff frequency of 15.9kHz. If C is ten times the size of C_R and the clock frequency remains at 100kHz, the cutoff frequency decreases to 1.59kHz.

Thus, the switched-capacitor filter has two significant **advantages** over the active (linear) one:

1. A low cutoff frequency can be achieved with capacitor values small enough to allow integration.

2. The cutoff frequency is not influenced by absolute variations. Given an accurate clock frequency and capacitor ratios of 1%, the cutoff frequency will be within 1%.

The simple low-pass filter can be expanded into any of the configuration discussed under active filters. Take for example the Sallen & Key filters in figure 13-7. In a switched-capacitor design you would first

greatly reduce the values of the capacitors and then replace the resistors with a capacitor and switches.

The switched-capacitor filter requires lateral switches, which are easily implemented in CMOS, but cumbersome (and slow) in a bipolar process. For this reason, this approach has become exclusively CMOS territory.

To minimize the influence of stray capacitances four (CMOS) switches are often used instead of two, resulting in an inverting configuration.

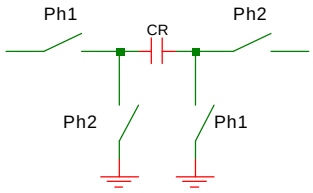


Fig. 13-22: Switch configuration to minimize the effect of stray capacitances in CMOS.

For either switch design it is important that the two lateral switches never be closed at the same time, i.e. there must be some "dead-time" between the two phases of the clock.

There are four **disadvantages** with switched-capacitor filters:

1. No matter how carefully you design the switches, there is always some switching noise.

2. A switched-capacitor filter *samples* the signal. To get an adequate sample, the highest signal frequency cannot exceed about 10% of the clock frequency. If there are signals present above that point, the switched-capacitor filter will produce a mixture of new frequencies, some of which may appear in the 0 to 10% frequency range. To avoid such false signals, a linear (active) filter must be used at the input (an **anti-aliasing filter**).

3. With an ordinary simulator, switched-capacitor filter can only be analyzed in real time; you cannot take advantage of the many features of an AC analysis, such as measuring frequency and phase response. And with the clock frequency necessarily being high, simulation takes far more time compared to an active filter. Only if the simulator has additional features, such as time delay in the AC model, can it give close to the same picture as that offered by linear AC analysis. There are some programs that have been designed exclusively for the analysis of switched capacitor filter.

4. The output has sampled noise, which is present even if the input is zero.

14 Power

Linear Regulators

Let's say you have 12 Volts available but need 3.3. Your 3.3-Volt load consumes up to 500mA. The 12-Volt source (e.g. a car battery) fluctuates between 10 and 14 Volts; the lower voltage needs to be within 5%.

The immediate choice to effect this change in voltage is a linear regulator. Look at it as a variable resistor, dropping whatever voltage is not needed.

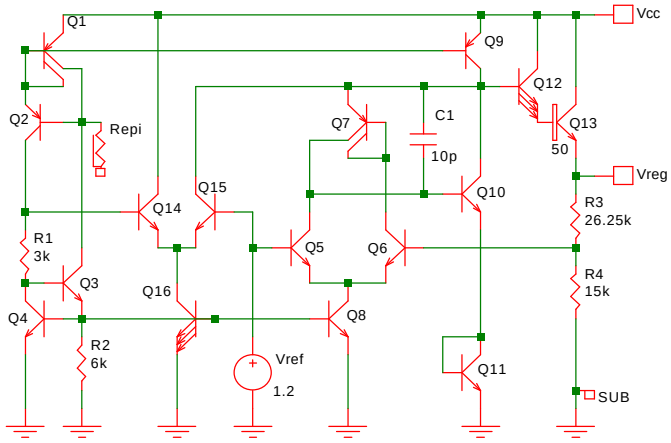


Fig. 14-1: Linear regulator with NPN power stage.

The unwanted voltage is dropped in an NPN transistor. In figure 14-1 this is a Darlington configuration to minimize the drive current; it requires at least 2.2 Volts difference between Vcc and Vreg, but it is an easy and simple design.

The regulator uses a 1.2-Volt bandgap reference (see chapter 7), whose voltage is compared with a fraction of the regulated output by the differential amplifier Q5, Q6, Q7 and Q10. Once the circuit is in balance the voltages at the bases of Q5 and Q6 are equal, so the regulated voltage is:

$$Vreg = \frac{Vref * (R3 + R4)}{R4}$$

An operating current is set up by Q1 to Q4 (a circuit derived from figure 5-4) and mirrored by Q9. At this point we have about 150uA and the current has a deliberate negative temperature coefficient (R2, which creates this current, is connected across a VBE, which itself has a negative tempco). This counteracts the positive tempco of hFE.

Q10 shunts to ground whatever operating current is not needed by the output stage.

Using a Darlington configuration for the output greatly reduces the required operating current, but there must always be a substantial voltage drop between supply and output. For this reason such a circuit is anything but a low-dropout regulator. For our application, a conversion from 10 Volts min. to 3.3 Volts this is of little concern.

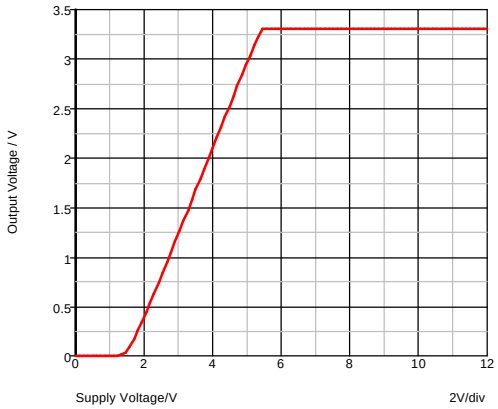


Fig. 14-2: Drop-out voltage of NPN regulator.

The current that flows through the load also flows through the output transistor. So, at 500mA, the load consumes 1.65 Watts, the regulator 4.36 Watts (with 12-Volts in), which is simply converted into heat. This the main disadvantage of a linear regulator. The heat is produced mainly by one device: Q13. Thus there will be a hot-spot on the chip and resulting temperature gradients, even with an adequate heat-sink. These temperature gradients are bound to influence other circuitry on the chip, including the regulator's own reference.

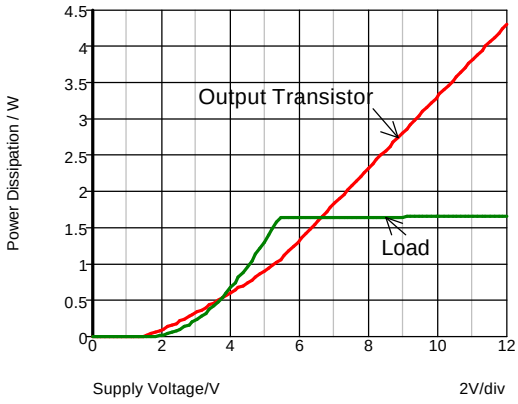


Fig. 14-3: In a linear regulator the energy not required by the load is converted into heat.

A linear regulator with an NPN output transistor is relatively easy to compensate. Despite the fact that the loop gain is high (which results in an output impedance of a mere 4mOhm) the circuit is rendered stable with a

single 10pF compensation capacitor. This stability holds even if a filter capacitor (of any size) is added at the output.

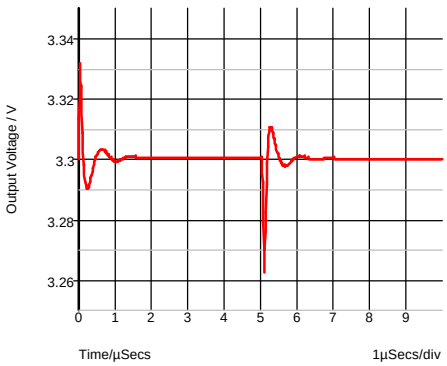


Fig. 14-4: The regulator is stable, even with a filter capacitor at the output.

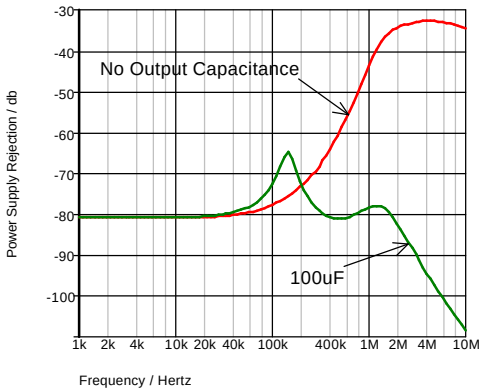


Fig. 14-5: Power supply rejection with and without a filter capacitor.

Because of the low output impedance it takes a massive capacitor at the output to have an effect of power supply rejection (figure 14-5).

There are three transistors and a resistor in this design which we have not discussed yet. The differential pair Q14/Q15 compares the reference voltage (which is assumed to have a very small temperature coefficient) with a voltage slightly higher than 2 VBE (which has a strong negative tempco). At temperatures below about 120°C the voltage at the base of Q14 is higher than Vref and Q15 is cut off. But at about 140°C these two voltages become equal and Q15 diverts the operating current for the output stage. Thus when the chip gets too hot, the output collapses and the source of the heat disappears. This makes the regulator virtually indestructible. As the Monte Carlo analysis indicates (figure 14-6) the accuracy of the shut-down point is ± 10°C.

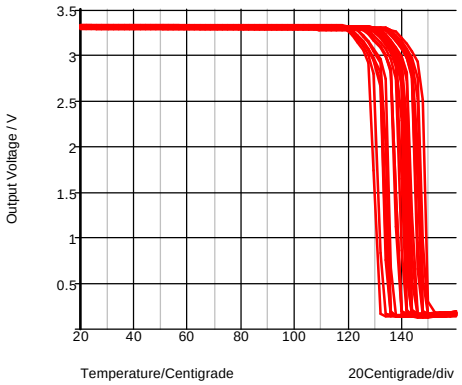


Fig. 14-6: Temperature shut-down.

Low Drop-Out Regulators

To get a lower minimum voltage drop we need to replace the NPN Darlington transistor at the output with a PNP (or P-Channel) device. And here is where the problem starts.

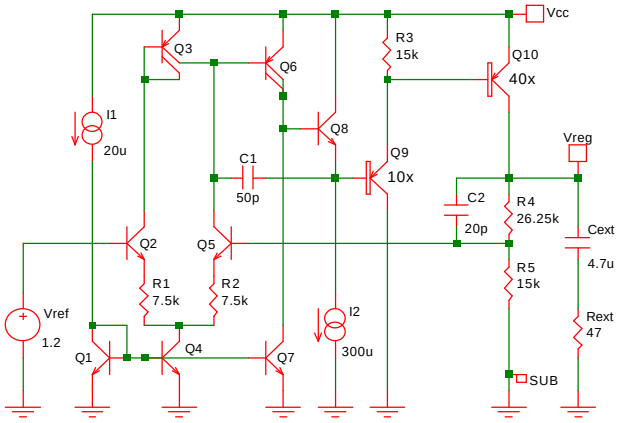


Fig. 14-7: Low drop-out regulator with internal (lateral) PNP transistor at the output.

Output transistors need to be large to carry the current and thus have substantial capacitance, multiplied by the Miller effect. This forms an additional pole, which gives the regulator a stubborn tendency to oscillate.

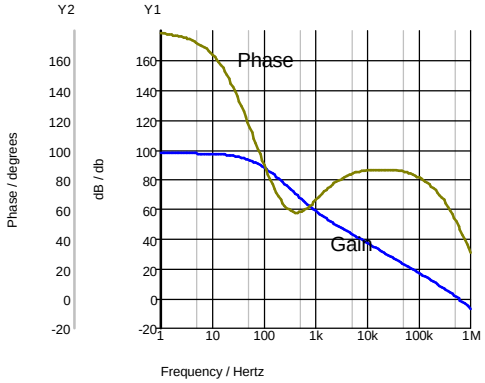


Fig. 14-8: Phase/gain diagram using three compensation capacitors.

It takes three capacitors to quiet down this regulator. C1 provides the main pole at about 30Hz. C2 corrects the phase at very high frequency and Cext, together with Rext form a zero at 1kHz. In addition the loop gain is reduced with R1 and R2. Even so, the phase margin is barely 50 degrees.

The external capacitor offers the benefit of increased power supply rejection, but its effectiveness is limited by the series resistor (which is essential to form the zero and turn the phase back up). At frequencies above about 5kHz the power supply noise appearing at the output is simply determined by the

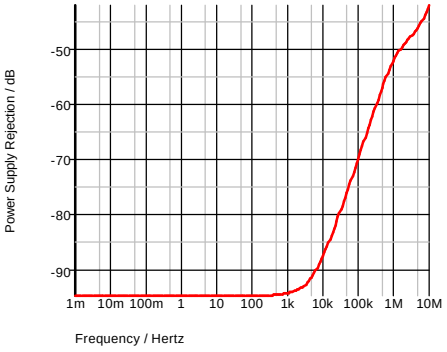


Fig. 14-9: Power supply rejection.

collector capacitance of Q10 and Rext.

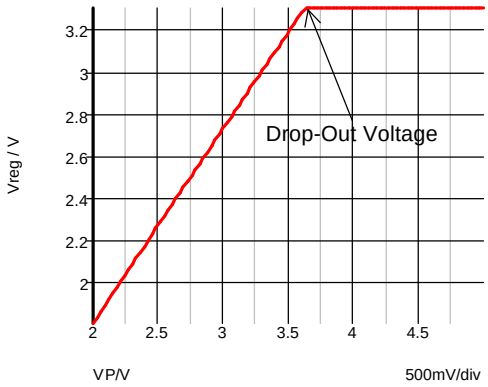


Fig. 14-10: Drop-out voltage at 20mA.

Q10 is a large lateral PNP transistor with an effective emitter length 40 times that of a small device, which makes it capable of carrying about 20mA. At this current the drop-out voltage is 300mV and the output impedance is 4mOhms.

Current capability can be increased at will, simply by making the output transistor larger. Low drop-out regulators using lateral PNP transistors have been built for up to 5 Amperes, but at such high current levels it helps having a special process which provides a higher doping level for the emitter. Even so, the output devices take up some 80% of the chip area. Be aware that, as the lateral PNP transistor saturates at the drop-out voltage, its substrate current becomes very large.

High Currents in an IC

There are two factors which limit how much current an IC can carry. The first is **electro-migration**. The force of huge numbers of electrons rushing through a conductor can become so large that the electrons begin to move atoms, physically push them along. For pure aluminum this happens at about 500'000 Amperes/cm^2. The effect is slow, it may take months, but eventually there will be an area where there is no aluminum left. Electro-migration is aggravated by high temperature and depends on the composition and grain structure of the metal.

Half a million Amperes may seem large and safe, but when you consider that you are dealing with very thin layers, the limitation becomes real. For example, for a thickness of 10'000 Angstroms (10'000Å = 1um) and a width of 1um a current density of 500'000A/cm^2 is reached with just 5mA.

The second limitation is resistance. Pure aluminum has a resistivity of 2.8uΩcm. Thus a layer 1um thick has a sheet resistance of 28mOhms/square. Make this run 100um long and you have a resistance of 2.8Ohms.

Let's say you want to carry 1 Ampere over a distance of 1000um on a chip. With a thickness of 1um, the aluminum stripe would have to be at least 200um wide to avoid electro-migration. It then would have a resistance of 140mOhms, i.e. drop 0.14 Volts.

And don't forget to check how much current contacts and vias can take in your process, as well as the thickness required for bonding wires.

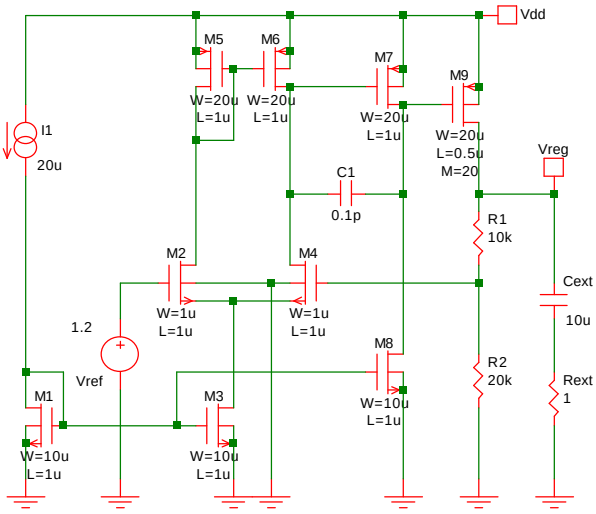


Fig. 14-11: Low-drop out CMOS regulator.

Figure 14-11 shows a CMOS version of figure 14-7. Also dimensioned for 20mA, the P-channel output device is smaller than the previous lateral PNP transistor. A low dropout voltage is, however only present at low current; to get the same value at 20mA, M9 would need to be 20 times the indicated size, or a total width of 8000um.

The circuit was designed for a supply voltage of 3 to 3.6V and a regulated output of 1.8V.

Frequency compensation is nearly as difficult as in the previous example. Again an external capacitor with a resistor in series is necessary at the output to create a zero and turn the phase up.

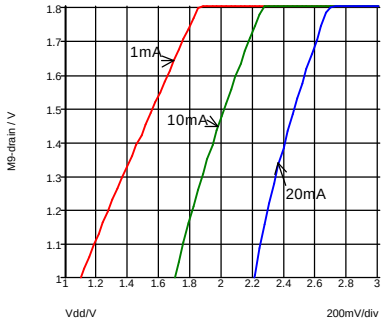


Fig. 14-12: Drop-out voltage.

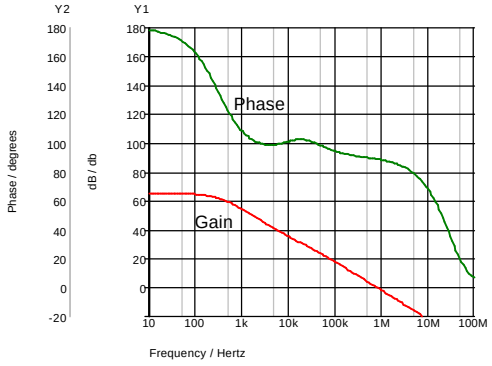


Fig. 14-13: An adequate phase margin is achieved with Rext ....

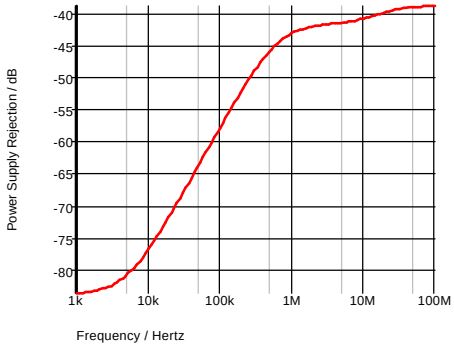


Fig. 14-14: ...but Rext limits power supply rejection at high frequency.

The last linear regulator makes the most sense for higher-current applications. Using an external PNP power transistor, it requires an extra pin, but it greatly reduces the area and power dissipation of the IC. Also, depending on the external transistor used, the drop-out voltage can remain low even at high current.

With the chosen device for Q6, the maximum current is around 500mA. At this level the supply voltage can drop to within 200mV of the output (3.3V).

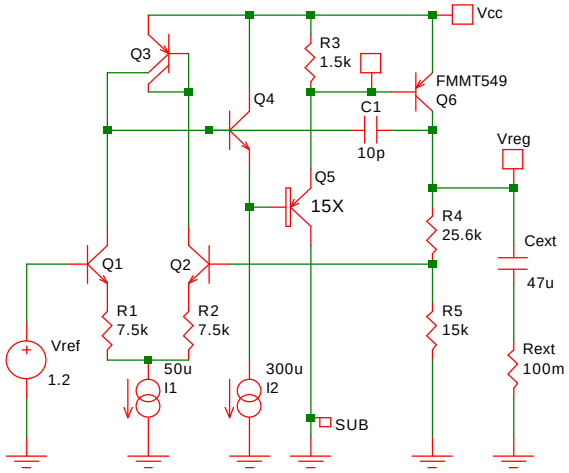


Fig. 14-15: Low drop-out regulator with external PNP transistor.

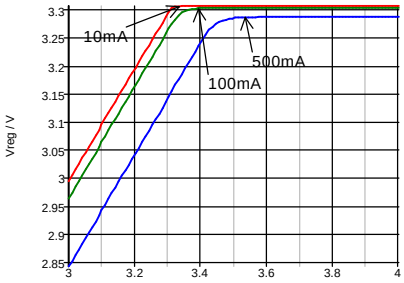


Fig. 14-16: Drop-out voltage.

There is, however, a compromise in the loop gain, which effects the output impedance (33mOhm); in order to achieve stability, the gain has to be reduced (R1, R2). Also, the same scheme as used in the two previous circuits is required: an output capacitor with a resistor in series to keep the phase from reaching zero before the gain does.

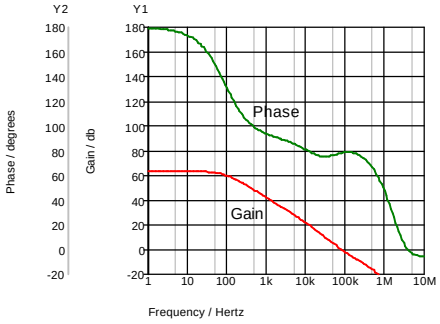


Fig. 14-17: Phase margin can only be kept high by a resistor in series with the output capacitor ....

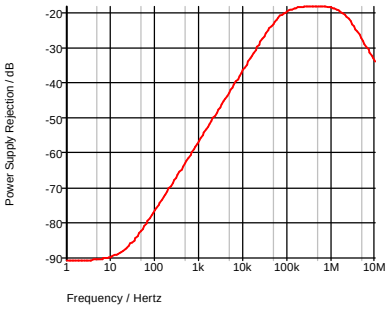


Fig. 14-18: ... which again impairs power supply rejection at high frequency.

Switching Regulators

Assume again that you have a supply voltage of 12 Volts, but you need 3.3 Volts. Your load consumes 1 Ampere.

A linear regulator acts as a resistor which drops the unneeded 8.7 Volts. In the process it converts 8.7 Watts into heat. 3.3 Watts are used by the load; a rather dismal efficiency.

Enter the switching regulator: instead of creating a resistance between input and output, it connects an inductor between the two for short periods of time.

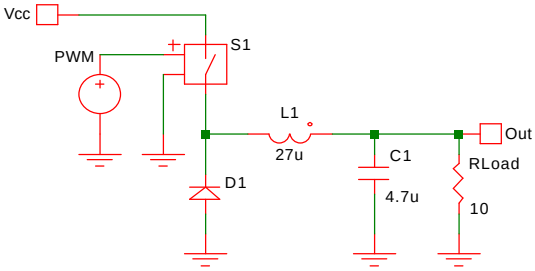


Fig. 14-19: Reducing a supply voltage with a series switch and inductor.

The switch, S1, is driven by a pulse generator (**PWM**, or **pulse-width modulator**). The pulses are rapid, so that the inductor value can be small.

The inductor, together with C1, smoothes out the switching pulses.

When the switch is closed, the left node of the inductor is at Vcc (assuming the switch has no resistance), but when the switch opens, this voltage jumps abruptly to a large negative value, created by the energy stored in the inductor. It is the purpose of D1 to catch this negative spike so it does no harm to the switch and provides a path for the current during the off period.

Figure 14-20 shows the resulting waveform at the output, for duty cycles of 10%, 20% and 40%. The average output voltage is simply proportional to the duty-cycle, but there is a noticeable ripple, the remains of the switching frequency (100kHz).

There is also an overshoot, which becomes more pronounced as the duty-cycle increases. This undesirable behavior is due to the LC filter (L1, C1).

With ideal components the voltage conversion is 100% efficient.

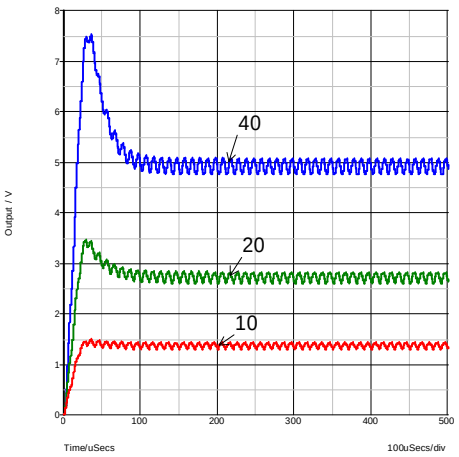


Fig. 14-20: The resulting waveform at the output.

But when you add some resistance to the switch and inductor and a forward voltage drop for the diode, the efficiency drops. For example, with a total resistance of just 50mΩ and a diode drop of 0.3Volts (a Schottky diode) the efficiency is 94%.

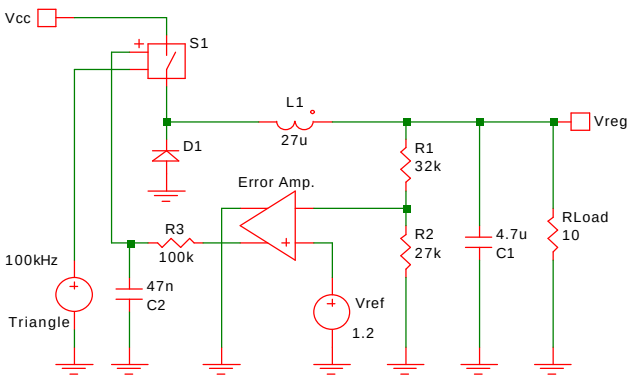


Fig. 14-21: "Buck" regulator.

The circuit of figure 14-19 is not a regulator; we have to add feedback to make the output voltage immune to supply fluctuations. This is accomplished by amplifying the difference between a fraction of the output voltage (R1, R2) and a reference voltage in an error amplifier. S1, an abstract simulation symbol, is now used as both a switch and a comparator (with the on/off thresholds set just a few millivolts apart). The output of the low-pass filter (R3, C2) following the error amplifier is thus compared with a triangle wave (100kHz, 2Vpp). In this way the regulator finds the duty cycle which gives the desired output voltage. Such a circuit is generally called a **Buck Regulator**.

There are a few items to consider, which are peculiar to a switching regulator:

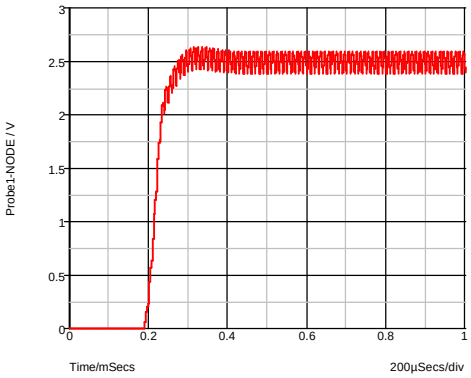


Fig. 14-22: Output voltage vs time.

First, an actual switch is not a perfect device; you will have to make a painful compromise, weighing voltage drop and speed: the lower the voltage drop the more current it takes to drive the device. For example, a discrete MOS transistor with an "on" resistance of 100mΩ at 1 Ampere has a total input capacitance of about 1nF. At a switching frequency of 100kHz you will need to turn the device on and off in less than 50nsec, otherwise the dissipation during switching becomes significant. This means the output of the comparator (the driver stage) has to provide 100mA to

charge and discharge 1nF. If you push the switching frequency to 500kHz, this current increases to 0.5 Amperes.

Second, the current level that the switching transistor needs to handle is always larger than the average output current. If you use a small inductor, the peak current can exceed the average by a factor of three or more; with a large inductance this factor is between 1.1 and 1.4.

Third, the voltage drop (and switching speed) of the diode is just as important as that of the switching transistor, their peak currents are roughly equal.

Fourth, the output LC filter (L1, C1) form a pole, which makes frequency compensation (R3, C2) more challenging.

We can step up the voltage by using the induced voltage in an inductor. Switch S1 connects the inductor L1 across the power supply (here assumed to be 1 Volt). The current flowing through the inductor is given by:

$$I = \frac{V * t}{L}$$

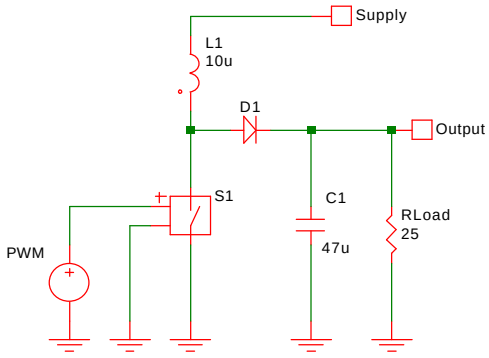


Fig. 14-23: By using inductive charge the output voltage can be made higher than the supply voltage.

As soon as the switch is turned off, a positive voltage appears at the anode of the diode, created by the stored current. This voltage is averaged by C1.

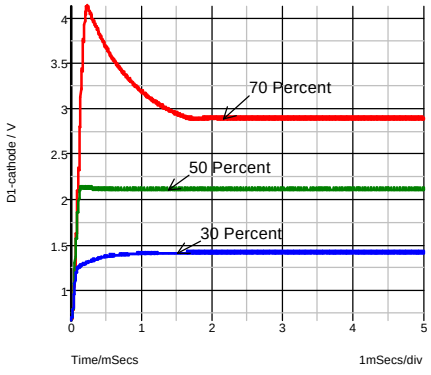


Fig. 14-24: Output voltage for three different duty cycles.

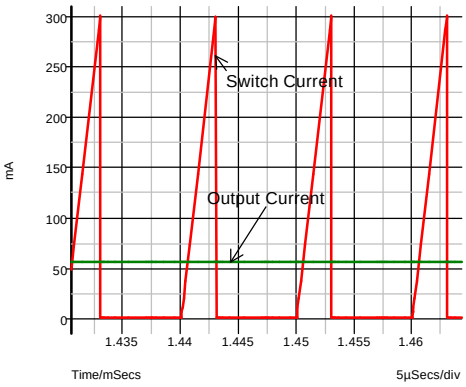


Fig. 14-25: Currents through switch and load.

The magnitude of the output voltage depends on how long the inductor is charged (i.e. what peak current is reached). Thus, by changing the duty cycle, the output voltage is altered. Note that in this configuration, too, the current the switching device must handle is considerably larger than the output current.

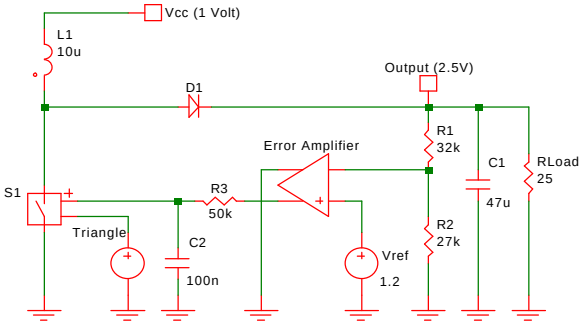


Fig. 14-26: Boost switching regulator.

Add feedback and we have a **Boost Regulator**. As before, the switch symbol represents both the switch and a comparator (i.e. the switch turns on and off within a few millivolts of the differential input signal). But be aware that, in this configuration, the feedback circuitry must have some specific characteristics: the output of the error amplifier must be constrained so that it stays within the amplitude of the triangle wave-form, otherwise the regulator can hang up at either zero or full output.

The frequency compensation network (R3, C2) also provides a "soft start", i.e. the output voltage builds up gradually, without much of an overshoot.

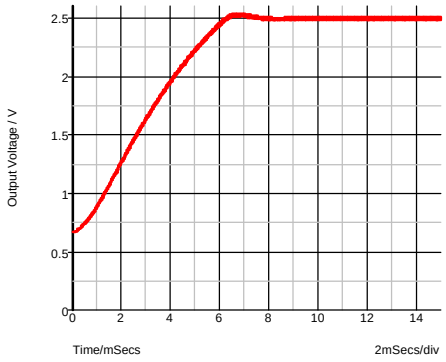


Fig. 14-27: Soft-start of the boost regulator.

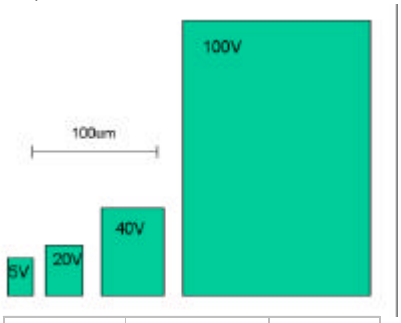
The same principle of using the inductive "kickback" voltage is also used to regulate larger voltages (such as a 110V or 220V line input). The inductor becomes a transformer, with a secondary winding delivering a lower voltage (isolated from the line). Feedback to the switching device is effected through an optical link (an LED and a phototransistor) to also provide isolation.

Some of the devices in such a line regulator, including the switching transistor, need to operate at high voltage; you need to be aware that this increases device size considerably (see panel).

The Voltage Penalty

As we have seen in chapter 1 (figure 1-15), depletion layers take up space. The higher the operating voltage, the wider the depletion layer. Thus the diffusions not only need to be deeper, but also more widely spaced.

Just how serious is the penalty of using large voltages in an IC? Take a look at the drawing below. It compares the required areas for minimum-geometry bipolar transistors operating at 5, 20, 40 and 100 Volts:



Max. Voltage	Dimensions, um	Area Ratio
5	22x29	1
20	30x37	1.7
40	52x70	5.7
100	152x220	52.4

If only a small portion of the circuitry is required to withstand a high voltage, you wouldn't want *all* of the devices to pay the price of large dimensions. This then calls for a more complex process, one capable of producing both shallow and narrow devices and deep and wide ones.

Linear Power Amplifiers

An ordinary amplification stage (e.g. figure 8-1) is categorized as **Class A**. There is a steady DC current through the transistor and, in the extreme, this current can be varied between zero and twice the idle value. The power efficiency of such a stage is dismal: it can only reach 50% at maximum output; with smaller signals it is much lower. In ordinary amplification we usually don't care about efficiency, but when it comes to a power output stage, class A is ill-suited.

In a **Class B** amplifier two output devices are used, one for the positive-going signal and one for the negative half. There is no idle current, each device starts to conduct as soon as the signal crosses the zero threshold.

This is an idealized concept which does not really work in practice. It is very difficult to switch from one device to the other without either leaving a gap or having both devices conduct at the same time. The result is distortion, which becomes very noticeable at low signal levels.

The solution is a compromise: allow a small idle current so that the amplifier works in a class A mode with small signals and *gradually* moves to class B as the signal increases. This operation is called **Class AB**.

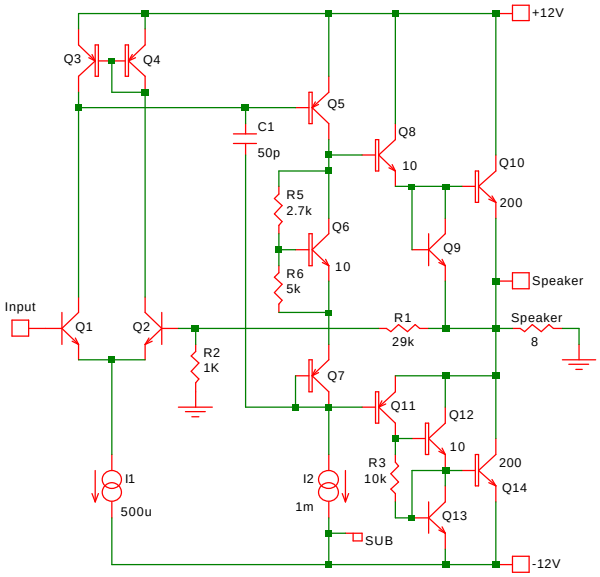


Fig. 14-29: 5-Watt bipolar class AB amplifier.

Such an amplifier is shown in figure 14-29. The two output devices are Q10 and Q14. They are large, having an effective emitter length some 200 times that of a minimum geometry transistor.

Ideally we would want one of the two output devices to be a PNP transistor, to exploit the complementary nature of the "push-pull" output. But NPN transistors carry a much higher current than PNP ones (unless a complementary process is available); with a 5.8 Watt output capability (requiring peak currents of 1.2A) this is no minor consideration.

To deliver the high output current, the upper stage (Q8, Q10) uses a Darlington configuration. Q9 serves to by-pass leakage current at high temperature.

The lower output stage has the identical Darlington connection plus a PNP transistor. The entire four-transistor block behaves like a PNP transistor. (All PNP transistors in this circuit are fairly large, capable of carrying 3mA).

There are three base-emitter junctions between the base of Q8 and the base of Q11. Between these two nodes a voltage is provided which causes a few hundred microamperes of idle current to flow through the two output transistors. This is done with the current I2 and transistors Q6 and Q7. The VBE of Q6 is increased with the resistor divider R5/R6 to the point where the desired current is reached. Notice that Q6 tracks the VBEs of Q8 and Q10 and Q7 tracks that of Q11.

The feedback resistors R1/R2 set the gain at 30dB and C1 provides frequency compensation. The slowest device in the amplifier is the compound PNP transistor Q11 to Q14, but it is fast enough to allow a more than sufficient frequency response for an audio amplifier without creating stability problems.

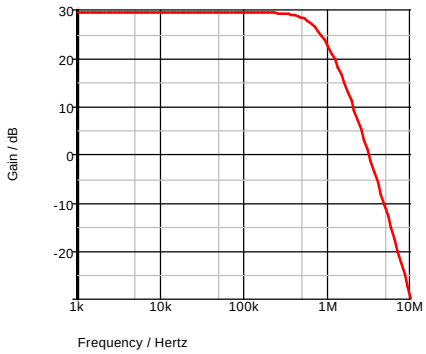


Fig. 14-30: Frequency response of the class AB amplifier.

One significant drawback of using only NPN power devices is voltage drop. Only ±10 Volts are available at the output from the ±12 Volt power supply without creating distortion. At 10Vp, however, the distortion amounts to only 0.15%.

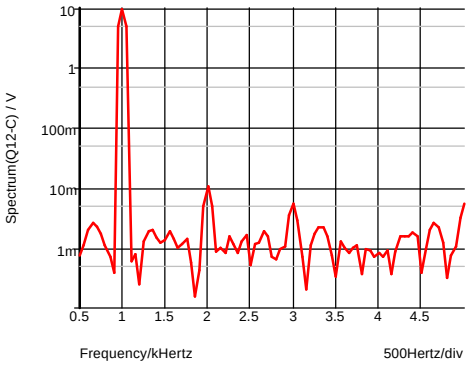


Fig. 14-31: Spectrum of the output signal at full power.

The maximum efficiency of an ideal Class B amplifier is 76%. For this circuit, with its 2-Volt drop in each output device, the maximum efficiency amounts to 62%. Thus the output transistors produce 1.7 Watts of heat each (for a 5.6 Watt output).

It is often argued that, in audio applications, peak power is rarely required and so the heat sink for the amplifier can be reduced in size. Unfortunately, in a class B (or AB) amplifier, peak dissipation occurs not at peak output, but at about 50% of maximum power.

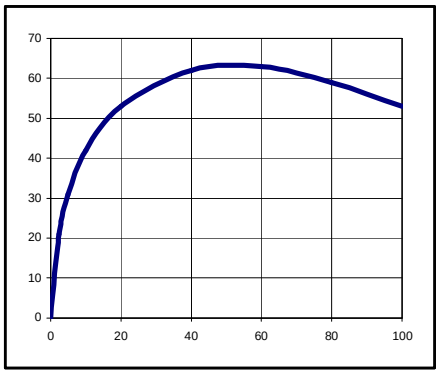


Figure 14-32: Normalized power dissipation (x) vs. power output (y) in a class B amplifier.

The design of figure 14-29 requires a split power supply. There are two ways to avoid this. We could convert the -12V connection to ground, make Vcc 24 Volts, bias the input at 1/2 Vcc and couple the speaker through a capacitor. The only

problem with this approach is the size of the new capacitor: 2000uF to get a 3dB drop-off at 10Hz.

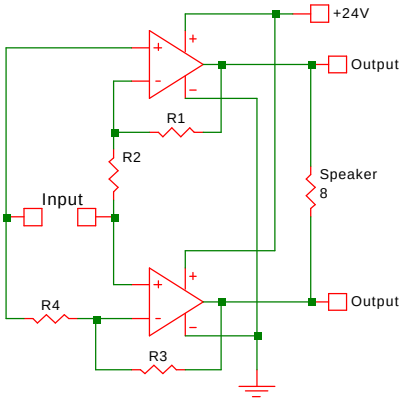


Fig. 14-33: Class AB amplifier with bridge output.

A better solution is the **Bridge Output**. In essence there are two amplifiers, 180 degrees out of phase. With no input signal, both output rest at 1/2 Vcc. As the signal appears, one output moves up, the other one down.

In this configuration we have in fact doubled the output swing. With the same total supply voltage, 25 Watts of output are generated (which requires *four* output transistors with a capability of 2.5A each). Efficiency is unchanged at 62%, which produces a power dissipation of 15.3 Watts.

Switching Power Amplifiers

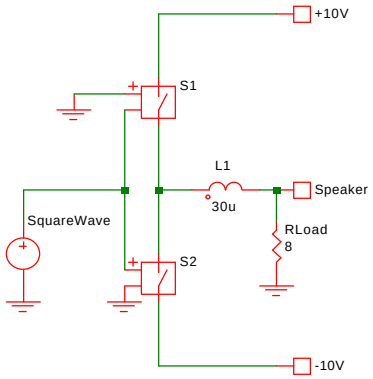


Fig. 14-34: Bidirectional switching arrangement.

The goal is almost the same as that of the series switching regulator: lower a voltage across a load without creating much heat. There are two differences though: the output starts at zero and it can move in either the positive or negative direction.

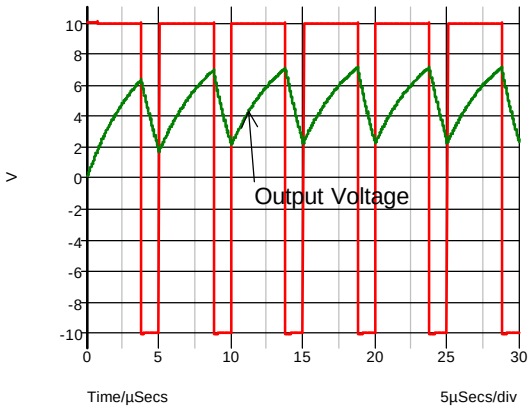


Fig. 14-35: Switching and output waveforms.

To start with, let's use two power supplies. The two switches connect the inductor to either the positive or negative rail. For now we assume that there is no dead

time or overlap and that this switching action is instantaneous.

The value of the inductor is fairly large for the chosen switching frequency (200kHz); it is never fully charged or fully discharged. Despite this, there is still a substantial ripple at the output.

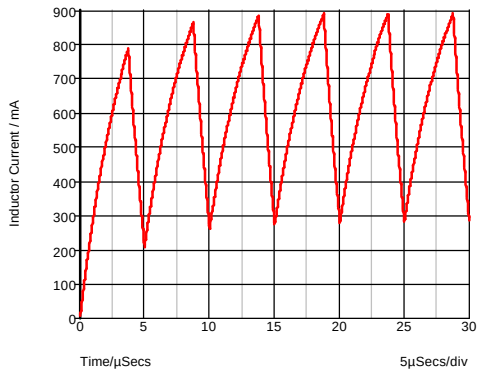


Fig. 14-36: Current through S1.

The average output voltage is a function of the duty cycle. At 50% the output is zero; 75% produces +5 Volts and 100% +10 Volts. Duty cycles of less than 50% cause the output to be negative.

Notice that the current gradually builds up (figure 14-36); the time constant of this effect is given by L1 and the 8-Ohm load (a speaker), a factor which will become important when we close the loop with feedback.

Let's now take the next step and modulate the duty cycle with a sine-wave signal, making a **Class D** amplifier. As in the switching regulators, the switch symbols also act as comparators (i.e. the thresholds of the control terminals are set so that the switches turn from off to on (and from on to off) within a few millivolts. Also (for now) the switches are ideal, they have no delay and insignificant resistance. Also, a filter capacitor (C1) has been added; this reduces the 200kHz ripple but increases the build-up delay mentioned above.

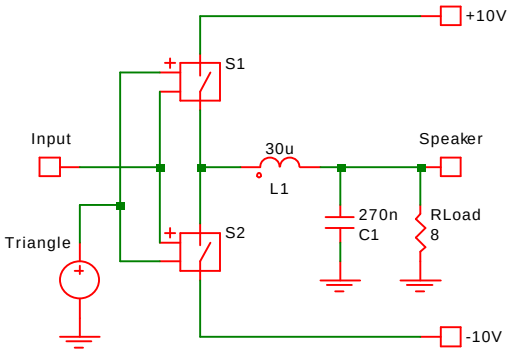


Fig. 14-37: Class D amplifier.

The output is now a sine-wave with a small amount of 200kHz ripple. Since we use near-perfect components, the distortion is very small.

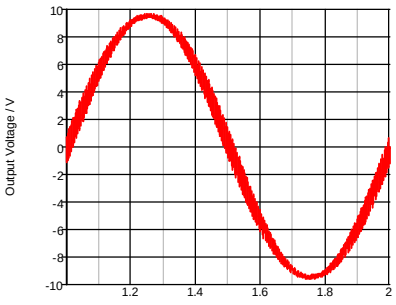


Fig. 14-38: Output wave-form.

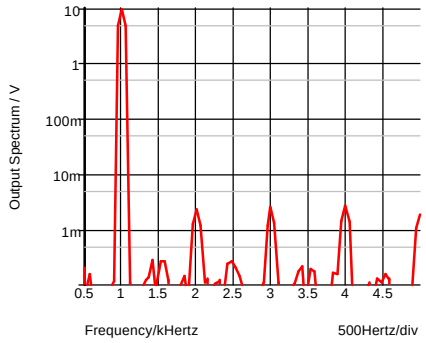


Fig. 14-39: Frequency spectrum in the signal range.

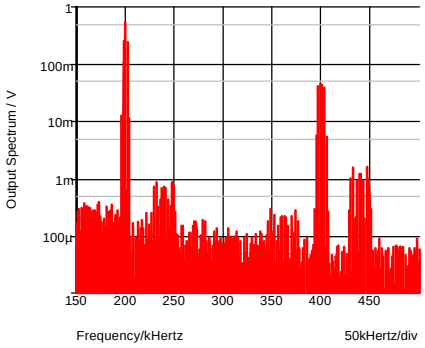


Fig. 14-40: Frequency spectrum in the switching range.

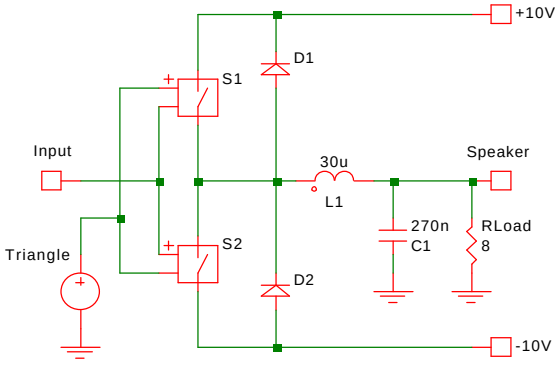


Fig. 14-41: Pulse-width modulated circuit with more practical component models. The diodes are now required to absorb the voltage spikes.

a P-channel one). In addition there is a small dead-time to avoid both devices being "on" at the same time. This dead-time creates a voltage spike from the inductor, which makes D1 and D2 necessary.

These small imperfections have a significant impact on the fidelity of the output signal: distortion increases to 1%.

Unless we use faster switching transistors with lower voltage drops and better matching, the

Alas, if we only had ideal components. In reality the switches have resistance and significant switching times. In addition, as pointed out on page 14-9, they require painfully large drive power.

In figure 14-41 the models are changed to represent more practical components. The switch resistances, for example, result in larger and unequal voltage drops (200mV for an N-channel transistor, 300mV for

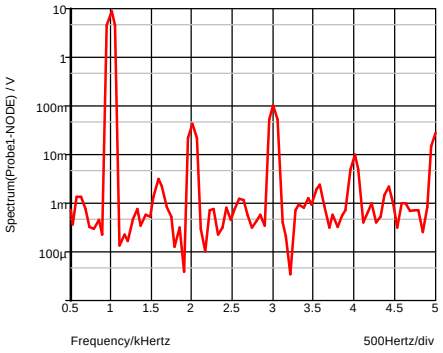


Fig. 14-42: Signal spectrum with realistic components.

level of distortion can only be brought down with feedback. And that is somewhat of a problem.

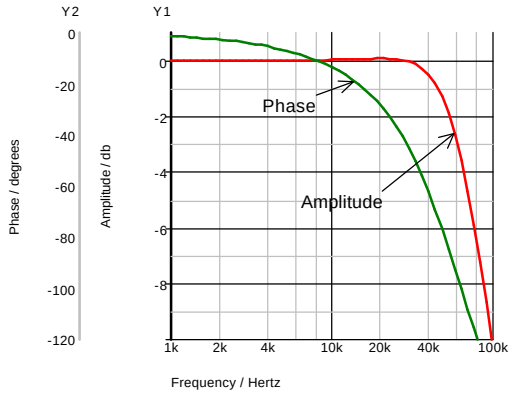


Fig. 14-43: Amplitude and phase response of the output filter.

In order to reduce the high-frequency components at the output we used an LC filter. It is dimensioned to be effective at 200kHz, but it causes a phase shift already in the audio range. Because of this, the amount of feedback possible, using a single feedback loop, is limited to about 20dB. With two or even three nested feedback loops this figure increases to about 35dB.

Also, a loud-speaker is not really a simple resistor, there is some inductance as well, making the phase relationship in the feedback path even more complicated.

We could, of course, increase the switching frequency, which would allow us to push the cutoff frequency of L1 and C1 higher, but the penalty would be lower efficiency and an increase in drive requirements for the switching transistors.

We have been assuming that we want a faithful (albeit larger) reproduction of the input signal at the load. Strictly speaking, this is not really true. In the case of an audio amplifier, the human ear cannot hear 200kHz, so filtering out high frequencies makes little difference. If the application is a servo amplifier, the load is unlikely to respond to such rapid fluctuations.

But there is radiation. Do we want to connect a square-wave of 200kHz (and its harmonics) across a long speaker cable and let it radiate into AM receivers and other electronic equipment? The answer is a clear no, and rules and regulations limiting such radiation have been written.

There are ways to reduce radiation. First, we can keep the speaker wires short, moving the amplifier next to the speaker. Second we can vary the switching frequency in a random fashion, creating a **spread spectrum**. Although this does not reduce the total radiation, it at least makes it less noticeable and allows meeting radiation limits.

For a given supply voltage and speaker impedance, the delivered power can be increased by using a bridge output. In figure 14-44 there are four switches. S1 and S4 are always "on" and "off" together, as are S2 and

S3. Thus the load is either connected to +V on the left side and -V on the right, or vice versa. This effectively doubles the supply voltage and the amplifier can deliver 25 Watts into an 8-Ohm load. There are four large output transistors, however, each of which must carry up to 2.5 Amperes.

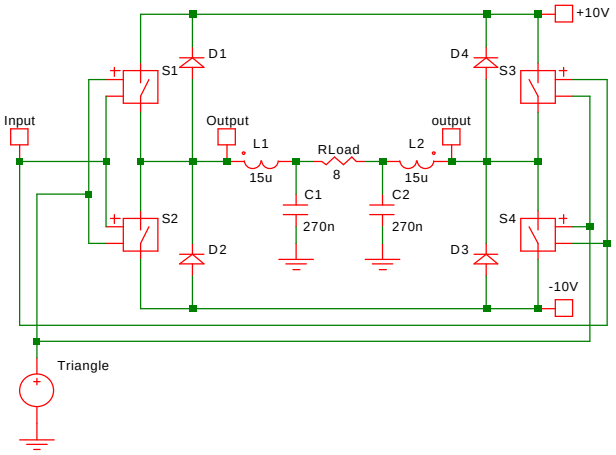


Fig. 14-44: Class D amplifier with bridge output.

If we apply 40 Volts total and use a 4-Ohm speaker, the output power grows to 196 Watts (and the peak current in the four output devices to 10 Amperes).

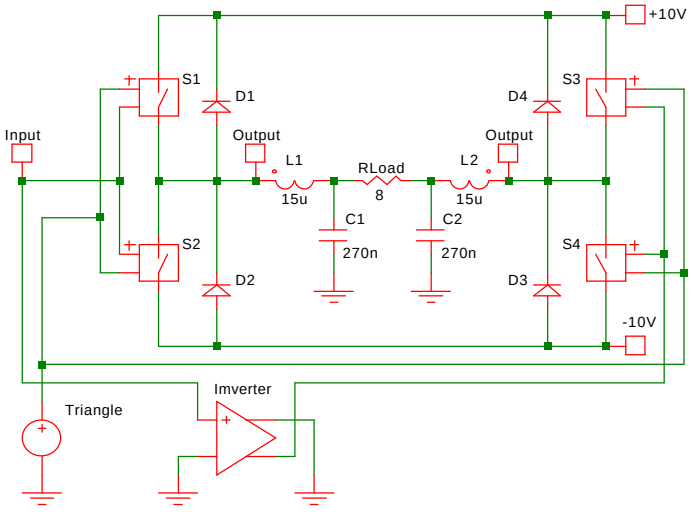


Fig. 14-45: Class D amplifier which suppresses the fundamental of the switching frequency.

But now let's change the circuit a little. Instead of having the two outputs move in opposite direction, invert one of the drives so that they move up an down together. If the input signal is zero, the two outputs will move at exactly the same time. Each output then carries a 200kHz square-wave, but *between them* there is no signal. As the input signal goes positive, the duty-cycle of one output increases while the duty-cycle of the other output decreases by the same amount. Thus, between the two outputs, there is now a square-wave with a duty cycle amounting to the *difference*.

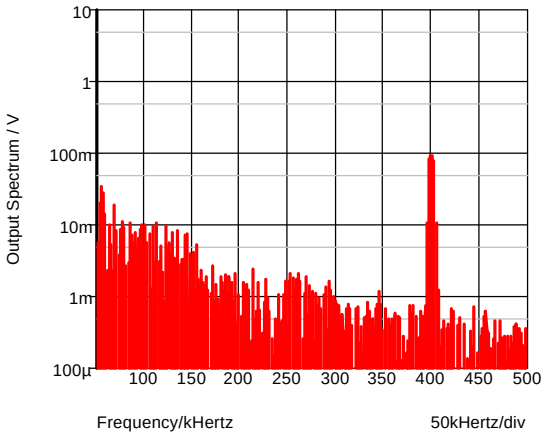


Fig. 14-46: Radiation spectrum across the load in figure 14-45.

The effect on the frequency spectrum is quite drastic: the fundamental of the switching frequency has disappeared; we only need to worry about the second harmonic, which has a lower amplitude and is easier to filter out.

But let's not get too enthusiastic here: the fundamental of the switching frequency is no longer present when measured across the load, but the wires leading to the load move up and down together, at the rate of the switching frequency. While this movement causes no current to flow through the load, there is still capacitive radiation from the wires. Hence C1 and C2 are needed.

A last word about class D amplifiers: simulation is very difficult and time-consuming. Unless you have a highly specialized program, only transient analysis can be used, which means you cannot obtain such parameters as phase margin directly. You may be forced to simulate (and integrate) the various blocks in pieces and then resort to old-fashioned breadboarding.

15 A to D and D to A

The field of data converters is vast and still expanding. It would be presumptuous to cover all of it in one chapter. For this reason only some of the most often used approaches are discussed here.

Digital to Analog Converters

There is nothing very mysterious about most digital to analog converters. Just take a look at the first figure. A string of identical resistors divides a reference voltage into eight equal parts. Of eight MOS transistors only one is on at a time, connecting the selected tap to the output.

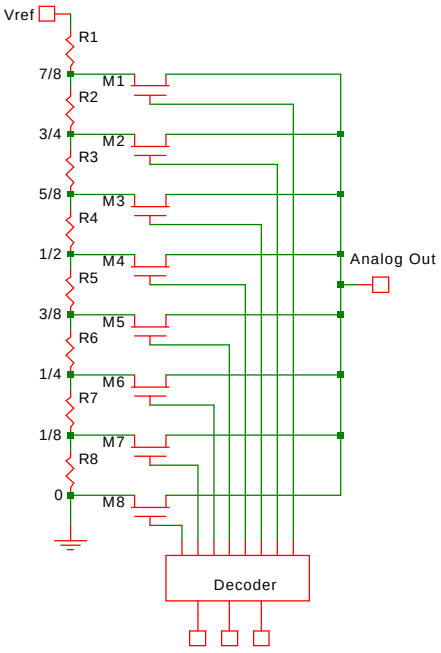


Fig. 15-1: 3-bit Divider DAC.

Of course this is a very simple example, a grand-total of three bits. It gets a bit more complicated as you increase the number of bits; 256 resistors and transistors are required for 8 bits, 1024 for 10 and 4096 for 12.

Also, this DAC creates analog voltages of only one polarity; analog signals have the nasty habit of being bipolar. To include negative-going values we need to double the number of resistors and add an identical negative reference voltage. Thus an 8-bit Divider DAC requires 512 resistor segments. For higher number of bits (and the expected higher accuracy) the matching of the resistors becomes the limiting factor; not only does the number of resistors increase but each resistor needs to be larger to obtain better matching. And forget trimming: at 12 bits you would have to trim each of the 8192

resistors.

Note that the full reference voltage does not get to the output. This quirk is caused by the fact that a string of eight resistors has nine nodes. We need to include zero, so three bits only reaches 7/8 of the total voltage.

To represent bipolar values some special codes are used. In the "sign + magnitude" code a bit is added which represents just the polarity. This is not the most efficient way and it is somewhat awkward (there are two values for zero, 0000 and 1000).

The offset binary code simply starts at the most negative number and counts up. Note that there is only one value for zero, but the full reference voltage is still not present.

In the twos complement code, positive numbers are the same as in a binary code, with the additional sign bit. Negative numbers are the inverse or complement of the positive ones, with a 1 added and the sign-bit changed.

Number	Sign + Magnitude	Offset Binary	Twos Complement
+7	0 1 1 1	1 1 1 1	0 1 1 1
+6	0 1 1 0	1 1 1 0	0 1 1 0
+5	0 1 0 1	1 1 0 1	0 1 0 1
+4	0 1 0 0	1 1 0 0	0 1 0 0
+3	0 0 1 1	1 0 1 1	0 0 1 1
+2	0 0 1 0	1 0 1 0	0 0 1 0
+1	0 0 0 1	1 0 0 1	0 0 0 1
0	0 0 0 0	1 0 0 0	0 0 0 0
-1	1 0 0 1	0 1 1 1	1 1 1 1
-2	1 0 1 0	0 1 1 0	1 1 1 0
-3	1 0 1 1	0 1 0 1	1 1 0 1
-4	1 1 0 0	0 1 0 0	1 1 0 0
-5	1 1 0 1	0 0 1 1	1 0 1 1
-6	1 1 1 0	0 0 1 0	1 0 1 0
-7	1 1 1 1	0 0 0 1	1 0 0 1
-8		0 0 0 0	1 0 0 0

Fig. 15-2: Codes representing bipolar values.

Other codes are also used for DACs. In the **BCD** (binary coded decimal) code each decimal digit is represented by four binary digits. BCD is primarily used for digital voltmeters. The **Gray Code** changes only one bit at a time, a feature useful in shaft encoders.

In any DAC, the output is strictly proportional to the reference voltage; double it and the output will double. Thus if you treat the reference terminal as an input, you have what is known as a **multiplying DAC**.

To increase the number of bits without exploding the number of resistors we can move to a **Segmented DAC**. Figure 15-3 shows a simple example for six bits, divided into three 2-bit segments. In this way we can reduce the number of resistors from 64 to 12; for a full-fledged design which delivers positive and negative values you would again need to double these numbers.

The first segment selects a resistor rather than a tap and the corresponding voltage drop is buffered and delivered to the second segment, where the same process is repeated. In the last segment taps are again

delivered to the output. Notice that in this approach, too, the top of R9 is not connected; a seventh bit would be required to allow that.

In such a segmented DAC the resistors in the first string (the most significant bits) are the most critical. They should be largest in size to obtain the best matching (or be trimmed).

It is crucial that a DAC be **monotonic**, i.e. as you step through the code from low to high, the output always increases (it may not increase by precisely the same amount, but at least it will never decrease). A divider DAC is always monotonic. The same holds true for the segmented DAC, provided each segment is monotonic (which is the case if we use dividers).

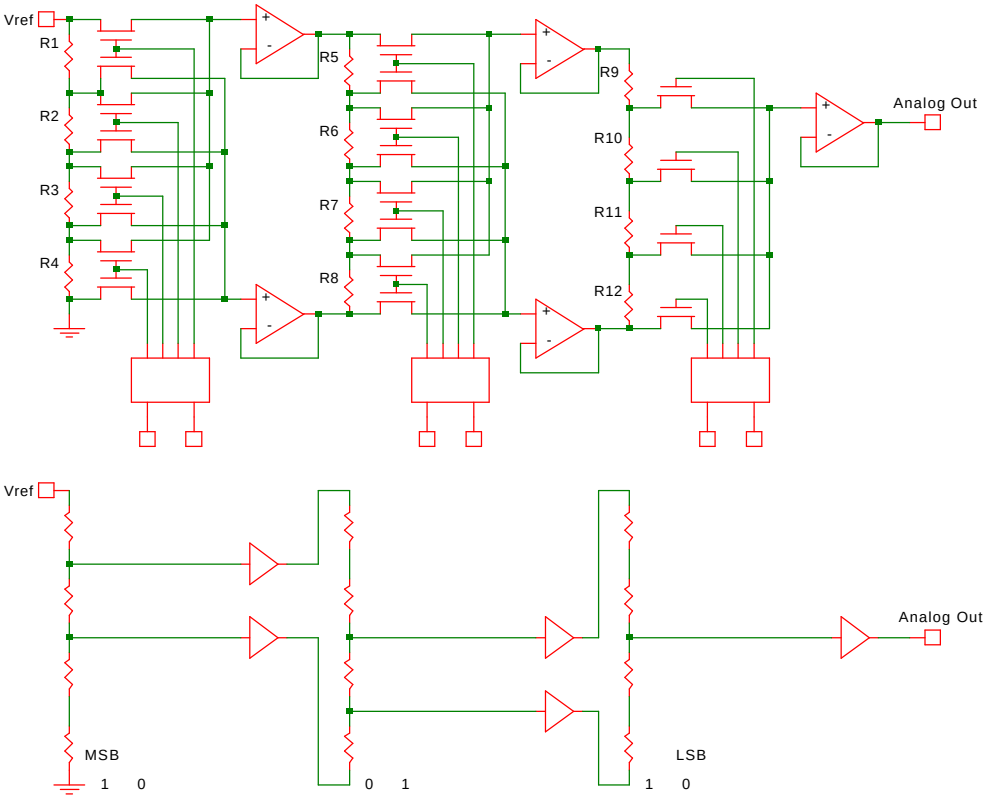


Fig. 15-3: 6-bit segmented DAC.

It is not necessary to use only identical resistors. Figure 15-4 shows a 6-bit example of a DAC using **binary-weighted** resistors. Moving from the most significant to the least significant bit, the resistors double in value each time, thus no decoding circuitry is required.

Only one resistor (and transistor) is required per bit but the saving is largely an illusion. To get good matching, you would need to design all

resistors with identical segments (1R), which amounts to 63 resistors, one less than a simple resistor string. In addition, the resistor for the most significant bit influences accuracy the most, so it should be larger than the others. Also, the transistors carry current, so their resistances appear in series with the binary weighted resistors. This means that M1 should not only be very large, but it should be 32 times the size of M6.

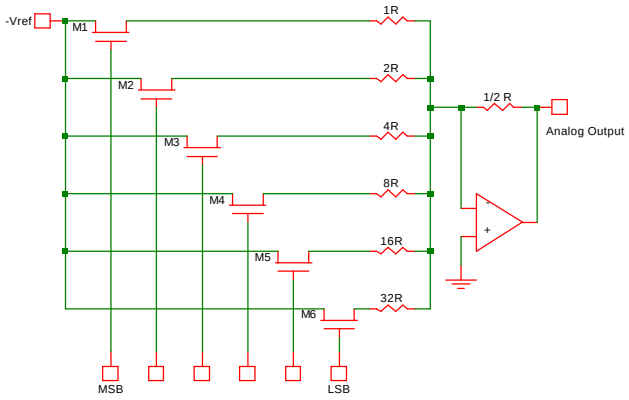


Fig. 15-4: DAC with binary weighted resistors.

A somewhat better idea is the **R-2R (Ladder) DAC**. Just two resistor values are used and the bit lines need not be decoded.

With the most-significant bit high, the first 2R resistor is connected between Vref and the input of the op-amp; all other 2R resistors are connected to ground. Each subsequent bit has half the influence on the output as the previous one.

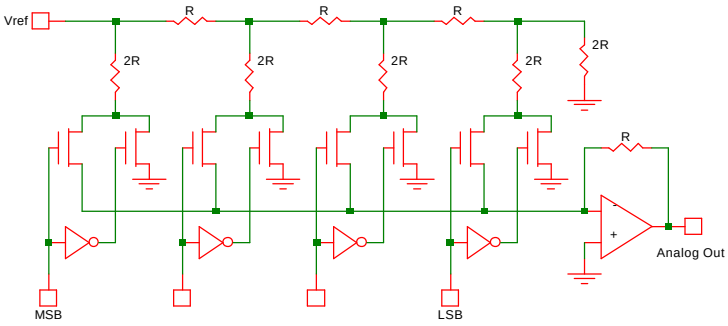


Fig. 15-5: DAC with R-2R ladder.

This is an inverting amplifier, so the output goes negative with a positive reference voltage.

Using only two resistor values improves matching and trimming is easier. Note, however, that the MOS transistors carry current and their resistance is critical. Moving from left to right, the current drops by 50% in each stage. Thus, to get the smallest error, the transistor size should be doubled for each stage moving from right to left.

In many DACs the analog voltage is created not by voltage taps but by currents. An example of a much simplified current DAC is shown in figure 15-6. A primary current is generated by R10 from a reference voltage; with Vref at 1 Volts, this current amounts to 200uA. Q1 through Q6, being biased from Q1, each produce a fraction of this current, Q2 100uA, Q3 50uA, Q4 25uA and Q5 12.5uA. Q6 is used to terminate the ladder.

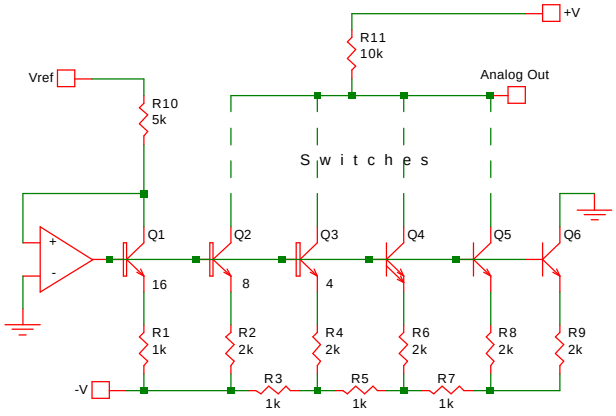


Fig. 15-6: Bipolar current DAC with R-2R ladder.

These binary-weighted currents are then switched to either R11 or +V (by, for example, a differential pair, acting as logic inputs). Note that the currents are flowing all the time, which makes this a fairly power-hungry approach.

You can, of course use the current directly as the output. But note that, if R10 and R11 are both inside the IC, their temperature coefficients and absolute variations will cancel.

Don't let this simplified example mislead you; there are many sources of error which require fine attention to detail. In a bipolar circuit there are base currents which must be compensated lest they subtract as much as 1% from the ideal values of the binary (collector) currents. And each collector must be at exactly the same potential as Q1 (here ground). Lastly, if bipolar switches are used, they, too, will have base currents which also must be rendered harmless.

Although it started out that way, a current DAC is no longer primarily a bipolar affair; in fact

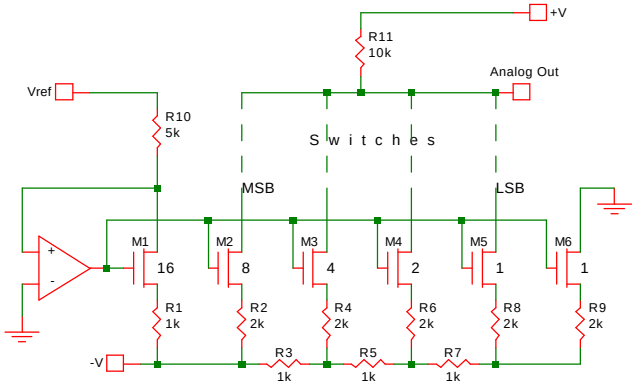


Fig. 15-7: CMOS current DAC with R-2R ladder.

CMOS has some significant advantages here. There are no base currents, therefore no base current errors. However, the drain voltages still need to be at identical levels, otherwise the Early effect will cause substantial deviation.

The number of bits is limited by the matching of the resistors and the sizes of the transistors. At eight bits M1 would consist of 512 transistors the size of M5 and M6. The latter two are already going to be far larger than minimum size to ameliorate the Early effect (making the channel long) and get acceptable matching (making the channel wide and long). 512 of these (twice as many for both polarities) make the area painfully large.

There is a way out of this limitation: segmentation. In our example

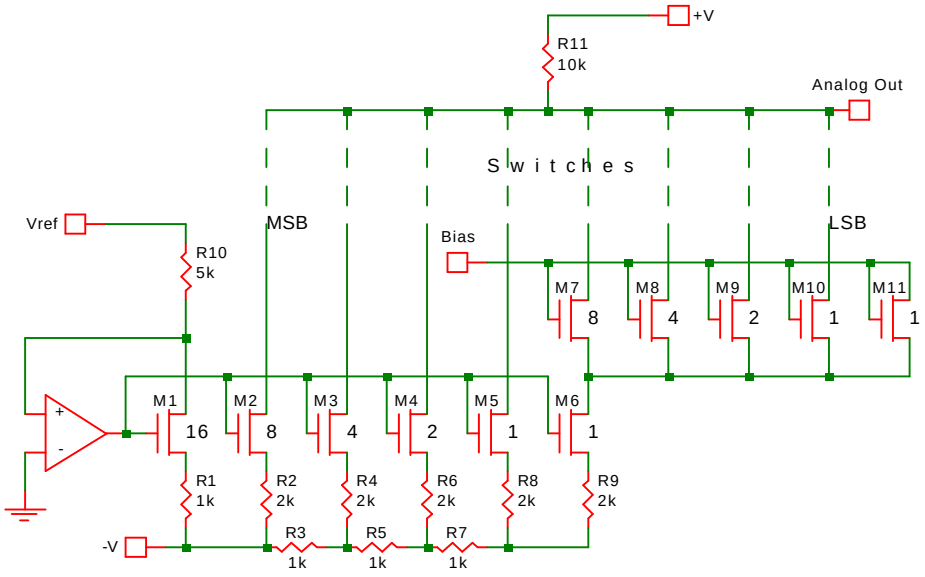


Fig. 15-8: Segmented current DAC.

the last transistor is only used for terminating the resistor ladder. It carries the same current as the least significant bit, 12.5uA. Use it and split it into 16 equal parts, 8 used for 5th bit, 4 for the 6th, 2 for the 7th and 1 each for the 8th bit and a dummy transistor.

Simple transistor ratios are used for this extension. We could have also employed another R-2R resistor ladder but, since these are the least significant bits, the accuracy is most likely sufficient.

Analog to Digital Converters

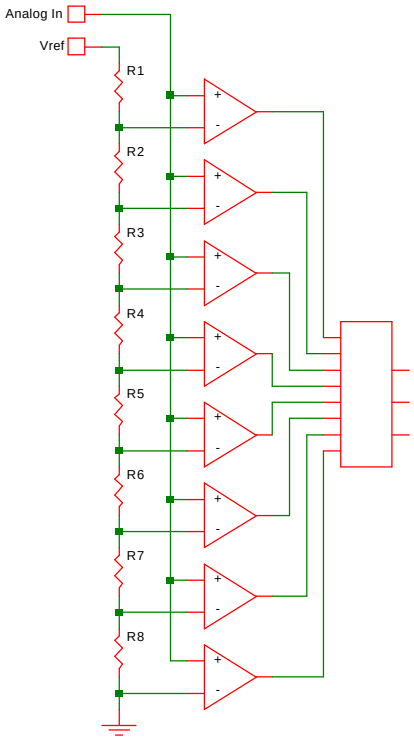


Fig. 15-9: Divider ADC.

As in DACs, the divider analog to digital converter is the fastest and most simple approach. Shown here for just three bits, there are eight comparators with one input connected to a resistor tap and the other to the analog input. Wherever the analog signal exceeds the potential at the tap, the output of the comparator goes high; the comparator outputs are decoded into three bits. If the input can be both positive and negative, twice as many resistors and comparators are needed, as well as a second reference voltage with an equal but negative value.

All comparators operate simultaneously, so the speed of the converter is given by the speed of the comparators alone.

The disadvantage of this approach is again complexity with large number of bits, with an accompanying high power consumption. At eight bits 512 resistors and comparators are required (assuming a bipolar input); at 12 bits the number increases to 8192. Also note that all comparator inputs are in parallel, which makes for a rather large input capacitance.

The **Successive Approximation** ADC reduces the number of comparators to one, though the number of resistors remains unchanged (Figure 15-10). Here a sample of the analog signal is taken and held steady while the conversion takes place. The heart of the ADC is a

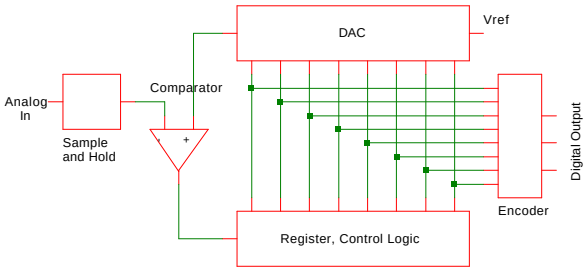


Fig. 15-10: Successive approximation ADC.

DAC. The control logic sets the DAC through a register to a likely value initially. If the value is too high, the register is moved down; if the initial guess is too low, the register is moved up. After a few steps the correct setting is found and the conversion stops.

All this guessing and stepping takes time, thus a successive approximation ADC is considerably slower than the divider approach, though it consumes less current and takes up a smaller area.

In both approaches the accuracy is limited by the resistor (or capacitor) divider, as was pointed out in the DAC section.

The Delta-Sigma Converter

Most engineers find explanations of the sigma-delta converter almost incomprehensible. There is a reason for this: terms are used which are quite non-descriptive and often misleading.

Take a look at a conventional diagram for a first-order delta-sigma ADC (Figure 15-11). This circuits has a "one-bit" output, which is more a riddle than a description Figure 15-12 shows the same function with more familiar blocks, which makes the circuit far easier to understand.

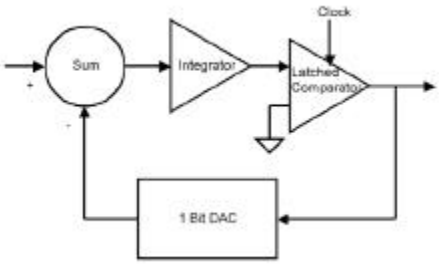


Fig. 15-11: Conventional diagram for a delta-sigma converter.

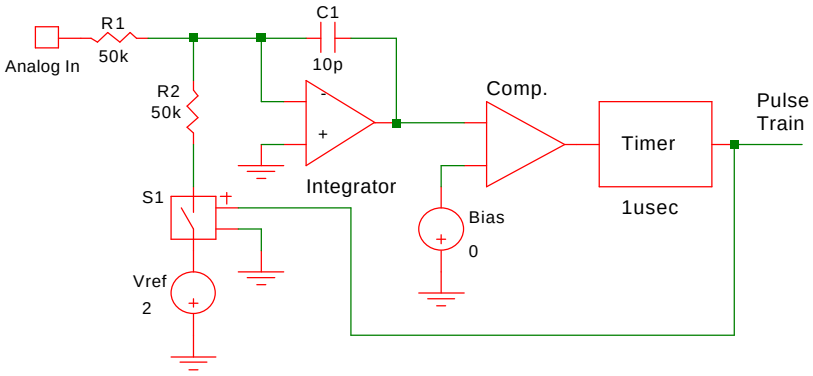


Fig. 15-2: First-order delta-sigma ADC with more familiar blocks.

The object is to produce a train of logic pulses at the output whose

frequency is proportional to the (analog) input voltage. Counting the number of pulses for a given time interval then gives the equivalent digital value, i.e. this is a voltage to frequency converter.

Let's start on the left. There are two resistors leading to the input of an integrator. R2 is connected through a switch to a negative reference voltage (of 2 Volts), while R1 responds to the input signal (with a range of 0 to 1 Volt). If switch S1 is open, the positive input voltage causes the output of the integrator to move negative at a rate given by R1 and C1 (this is an inverting integrator).

The following comparator then triggers a timer when this falling voltage reaches the bias level (which is set here at zero volts but can be any convenient level). The timer produces a short pulse (e.g. 1usec) which is delivered to the output.

This pulse also closes the switch. Since the two resistors are equal in value (an arbitrary choice) and Vref is at least twice the value of the input voltage, the integrator reverses during this time and its output moves up.

If the input voltage is 1 Volt, the output of the integrator moves positive during the pulse by exactly the same amount as it has moved negative while the switch was off. Thus the duty cycle is 50% and the frequency 500kHz (remember that R1 is connected all the time, while R2 is connected only half the time, hence Vref needs to be at twice the level of the maximum input signal).

If we lower the input signal to 100mV, the falling portion of the integrator output becomes longer (while the rising portion remains constant) and the frequency drops to 50kHz. At 10mV input the frequency is 5kHz and at 1mV 500Hz. At zero input the oscillation stops entirely.

As in the examples above, this circuit can only handle positive input voltages. For a bipolar input (say ± 1 Volt), R2 is switched between two reference voltages, +2V and -2V.

So, the mysterious "1 Bit DAC" turns out to be nothing more than a switch and one or two reference voltages, the "Latched Comparator" a timer triggered by a comparator and the "Summer" two resistors (it actually is a subtractor). The "delta-sigma modulator" (which purists insist should not be called a sigma-delta modulator) is simply a voltage to frequency converter. And the "1 Bit Output" translates into *serial output*.

For accuracy two factors are of overwhelming importance: the reference voltage and the pulse-width. Reference voltages can be trimmed and the pulse-width can be derived from a crystal-controlled clock. All other elements are of secondary importance. R1 and R2 need to match well, but their absolute values (and that of C1) only affect the height of the triangle wave at the output of the integrator, not the timing (the rising and

falling flanks of the triangle wave are equally affected and thus cancel out). A high loop gain in the integrator op-amp assures that the voltage fluctuation at its input is down to a few microvolts (but its offset voltage still matters).

The performance of a delta-sigma ADC can be improved by adding one (or more) feedback loops (Figure 15-13, shown again in the conventional way). Be aware, however, that the higher the order, the less stable the design becomes. A third-order delta-sigma ADC can oscillate in some unexpected ways.

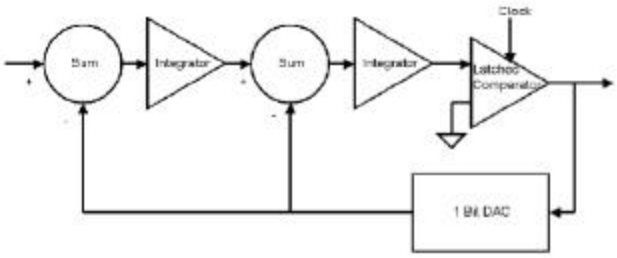


Fig. 15-13: Second-order Delta-Sigma ADC.

The significant advantage of the delta-sigma ADC is its capacity for resolution. These circuits are often called "oversampling ADCs", because they must sample the incoming waveform above the Nyquist rate (twice the maximum input frequency). For example, if you want to capture a 1kHz signal with an 8-bit resolution, the maximum frequency must be at least 2kHz times 256, or 512kHz. At 12 bits this frequency increases to 8.2MHz.

However, this presumes that we do nothing else than counting pulses at the output, which ignores much of the concept's capability. The delta-sigma ADC was made for CMOS and with today's small geometries a great deal of digital signal processing can be done once the pulses exist. Apart from increasing the resolution, the sampling noise (which is already centered around a rather high clock frequency) can be brought down to stunningly low levels with a sophisticated digital low-pass filter (called a "decimation" filter).

With these additional measures, a second-order delta-sigma ADC with a signal bandwidth of 4kHz can achieve a 14-bit resolution with 85dB signal to noise ratio, using a clock frequency of 1MHz.

16 Odds and Ends

In this, the last circuit design chapter, we look at six functions which did not fit well into the previous subjects.

As pointed out before, be aware that you will need to re-simulate these circuits with models specific to the process to be used.

The Gilbert Cell

An unusual and brilliant idea: Take two current mirrors and connect them in a differential way. Run the inner pair at a higher current than the diode-connected transistors and you get gain, roughly in the ratio of the currents. I2 and I3 are modulated, the collector currents of Q2 and Q3 are the outputs.

Since this is a current in/current out scheme, the cell is fast (no Miller effect).

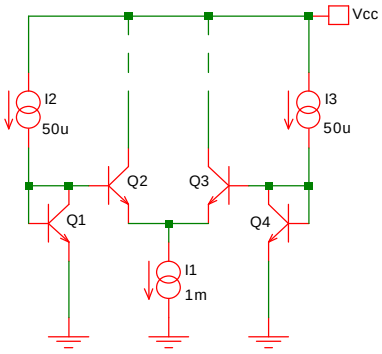


Fig. 16-1: The first form of a Gilbert cell.

In the second form of the Gilbert cell all three currents flow to the negative rail, which allows stacking: use the collector currents as the inputs for the next cell. Each subsequent cell is biased at a higher DC potential to avoid saturating any of the transistors.

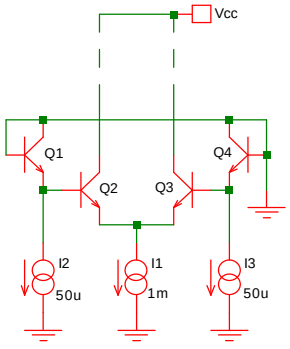


Fig. 16-2: Second form of the Gilbert cell.

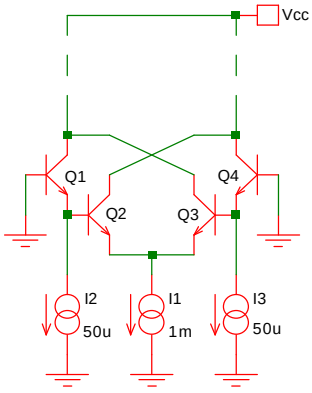


Fig. 16-3: And the third form of the Gilbert cell.

In both forms there is a small error due to the base

currents, which is largely eliminated in the third form.

Alas, all this is true only in an isolated, theoretical analysis. It is very rare that you start out with a differential current input. In most applications there is an input *voltage*, and single-ended at that. So, to use the Gilbert cell, you need to convert this voltage into a differential current, for example with a differential pair, as is shown in figure 16-4.

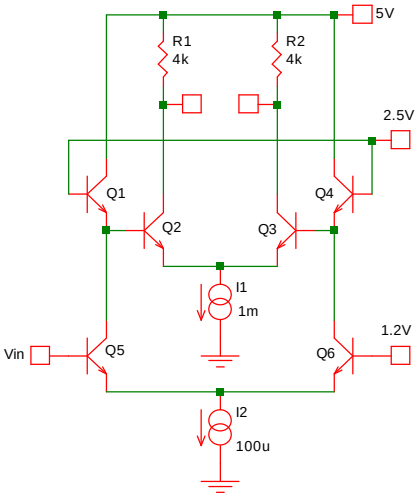


Fig. 16-4: A practical application of the Gilbert cell.

This problem is made worse by stacking several cells, requiring a wide range of current levels. In today's power-conscious and low-voltage environment the Gilbert cell has become outdated.

And here is where the Gilbert cell falls down. As shown in figure 16-5, a differential pair actually has a higher gain and wider frequency response (if operated at I1) *without* the Gilbert cell. Which is why it is rarely used. This is not due to the fact that the current is converted into a voltage at the output (and thus the Miller effect is right back in the picture), but simply because Q1 and Q4 need to run at a lower current.

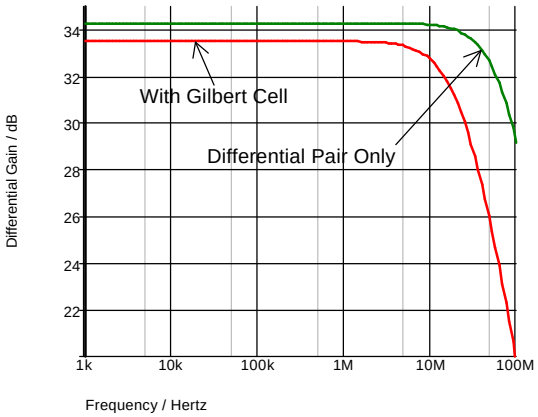


Fig. 16-5: In most applications the Gilbert cell does not actually enhance performance.

Multipliers

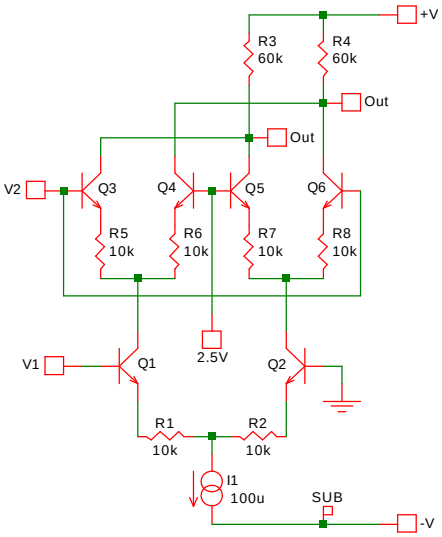


Fig. 16-6: Simple four-quadrant multiplier.

We have seen a similar circuit before, used as a phase detector for a PLL. While accuracy in that application was of minor importance, in a multiplier it is the main feature,

The circuit requires a split power supply (e.g. ± 5 Volts), so that at least one input (V1) can be at ground level. The second input (V2) is biased safely higher (2.5V) to avoid saturating Q1 and Q2.

It is the insertion of resistors in the emitters of all six transistors that gives this multiplier its accuracy. Their values need to be large compared to the dynamic emitter resistance (r_e, see page 4-1).

Such a circuit is called a four-quadrant multiplier because it produces an output for all four quadrants of a plot: 1. both inputs positive; 2. V1 positive, V2 negative; 3. both inputs negative and 4. V1 negative, V2 positive (where the value of V2 is applied with respect to 2.5V),

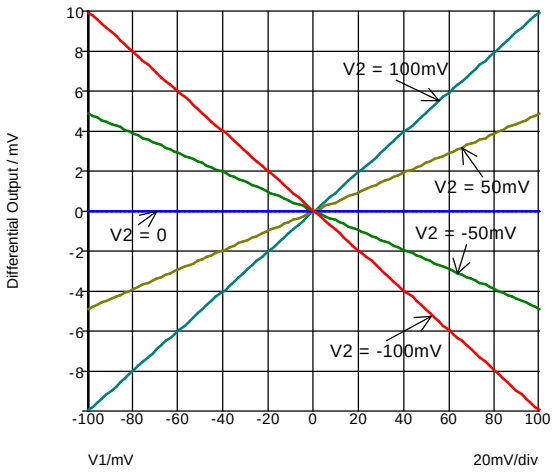


Fig. 16-7: Behavior of the four-quadrant multiplier.

The range of the two input voltages is ± 100mV, resulting in a maximum output of ± 10mV. This limits the achievable accuracy since matching of VBEs becomes as important a factor as matching of the resistors. A higher positive supply voltage is needed to allow input and output ranges of ± 1 Volt.

As shown, the error can be as high as ± 5% untrimmed and ± 1% trimmed. With a higher supply and trimmed thin-film resistors ± 0.3% is possible.

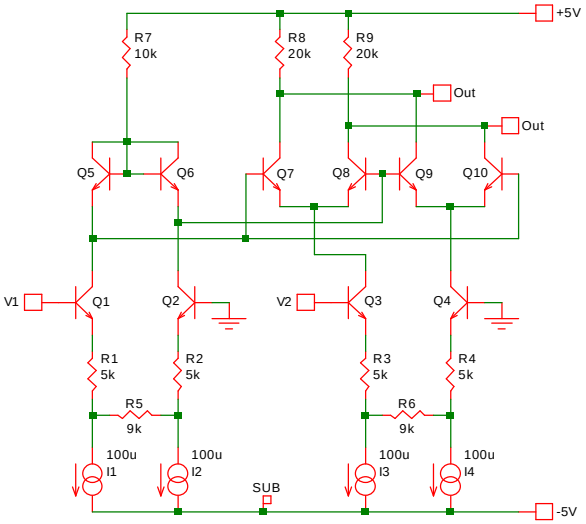


Fig. 16-8: Four-quadrant multiplier with both input voltages at ground level.

By adding two transistors and biasing the upper quad (Q7 to Q10) with diode-connected transistors (Q5, Q6, sitting at a DC potential set by R7), both input can be at ground level. Also, the ranges of the two inputs and the output are now extended to ± 1 Volt. Accuracy is unchanged for untrimmed operation; with trimmed thin-film resistors and additional temperature compensation (see reference) such a circuit can be brought to within 0.1%.

Figure 16-9 shows the equivalent circuit in CMOS, designed for a 0.35u process. Because of the lower supply voltages the range of the input voltages is again limited to ± 100mV.

This circuit illustrates the performance limitations imposed by low supply voltages. With offset voltages generally being higher in CMOS and the maximum output range a mere ± 10mV, untrimmed accuracy is no better than about ± 10%. You

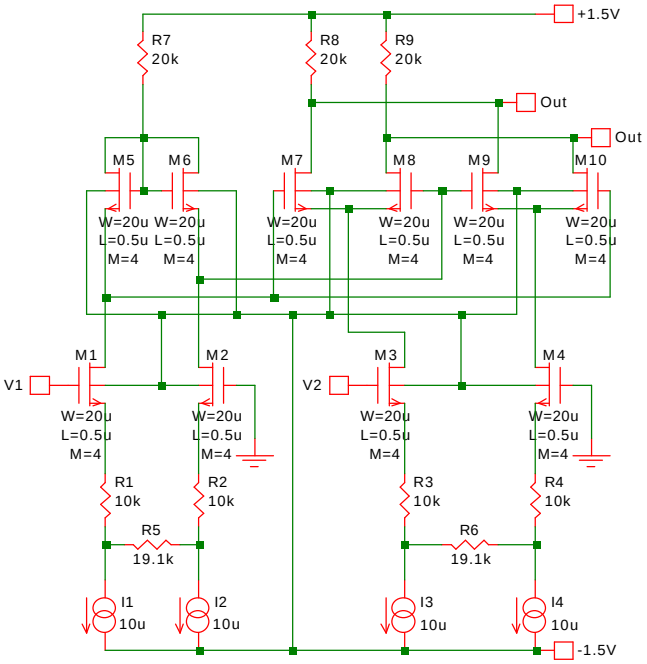


Fig. 16-9: Four-quadrant CMOS multiplier.

can, of course, change the resistor ratios so that the relationship between inputs and output is multiplied by a constant.

Peak Detectors

Peak detectors tend to be a bit tricky. A surprising number of the schemes using an op-amp and a diode tend toward oscillation or other misbehavior.

When you analyze the feedback loop, you find that the diode and the capacitor at the output place an unusual burden on the op-amp. When the signal moves up (for a positive peak detector) the output is connected to a large capacitance. When the signal moves down, there is no load at all.

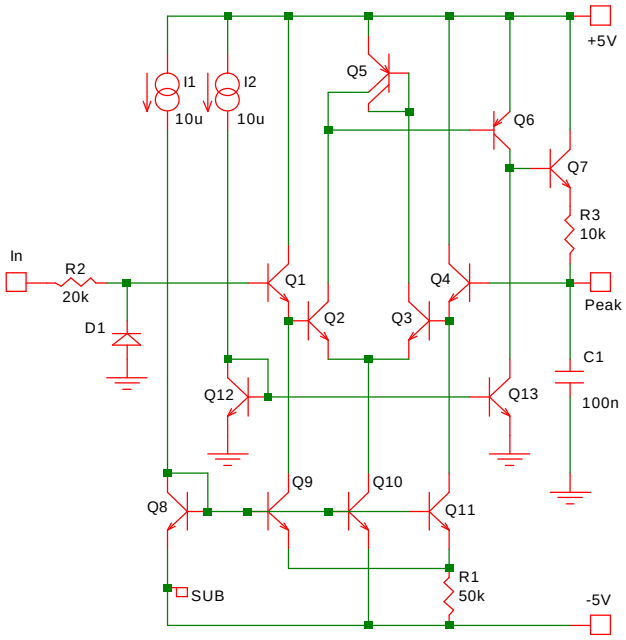


Fig. 16-10: Peak detector with Differential NPN Darlington pair.

The circuit in figure 16-10 uses a bipolar Darlington pair to provide a high-impedance input at the (external) capacitor. For operation over a wide temperature range the outer transistors (Q1, Q4) are biased at about 0.8uA by Q9 and Q11.

The output impedance of the op amp is artificially enlarged by R3 to provide frequency compensation (together with C1).

There is a fundamental question to every peak detector: How long should the voltage stay on the capacitor? If the answer is "forever" then the detector displays the highest peak for an infinite history of input signals, which is probably not what you want. For any other answer you have two choices: either discharge C1 slowly, so the voltage stays within the desired accuracy over the time of interest, or reset C1 before each measurement.

In our circuit the capacitor is discharged by the base current of Q4. This current amounts to about 5nA. If the peak voltage is 1 Volt, you lose about 1% in 200msec with 100nF of capacitance. The discharge current varies from chip to chip and with temperature.

Figure 16-11 shows a peak detector which operates from a single supply voltage. The Darlington input pair of the op-amp is more elaborate and the operating currents of the outer transistors (Q1, Q6) are merely the base currents of the inner ones. This limits the temperature range (to about 100°C) but lowers the input current.

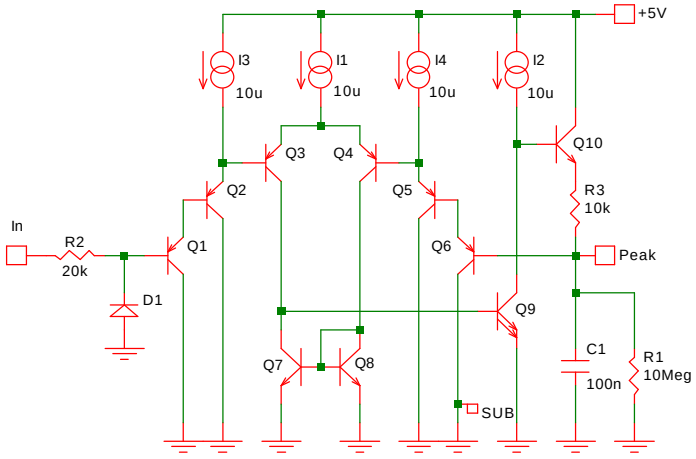


Fig. 16-11: Single-supply peak detector with a lower input current.

Notice the discharge resistor R1. Without it, the base current of Q6 would *charge* C1 (it flows out of the base). Thus, in this circuit, a controlled rate of discharge through R1 is essential.

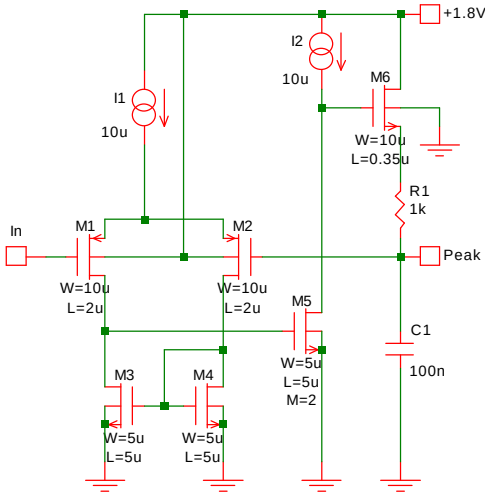


Fig. 16-12: CMOS peak detector.

In both examples the supply voltage must be lower than the emitter-base breakdown voltage of the output transistor. If it is to be higher, use an additional diode in series with the emitter.

CMOS devices are much better suited for peak detector design than bipolar ones for two reasons: 1) there is no (DC) input current and 2) you can reset a capacitor to zero volts (the collector-emitter voltage of bipolar transistors does not go to zero, there is always a remaining voltage of about 100 or 150mV).

As in the bipolar examples,

the feedback loop is compensated with a resistor in the output path, working together with Cext.

Since there is no input current, C1 can be made quite small, to the point where it can be internal. But be aware that the smaller C1 the more difficult it becomes to compensate the feedback loop.

Rectifiers and Averaging Circuits

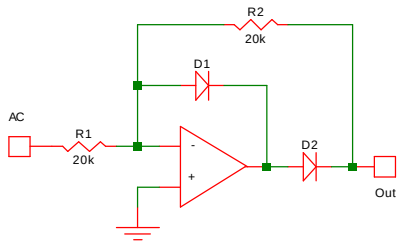


Fig. 16-13: Standard op-amp half-wave rectifier.

Figure 16-13 shows the standard configuration for a half-wave rectifier, appearing without much comment in dozens of text-books, usually without mentioning that it does not work well with many op-amps.

Putting a diode in the feedback path is awfully hard on the op-amp. The abrupt impedance change around zero signal level can easily cause spikes and damped oscillation, affecting the accuracy. Specifically designed circuits avoid this - and require only a single supply voltage.

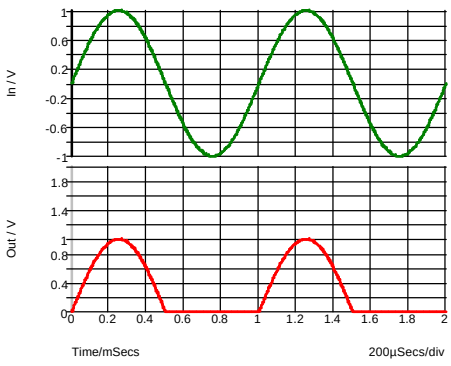


Fig. 16-15: Input and output waveforms.

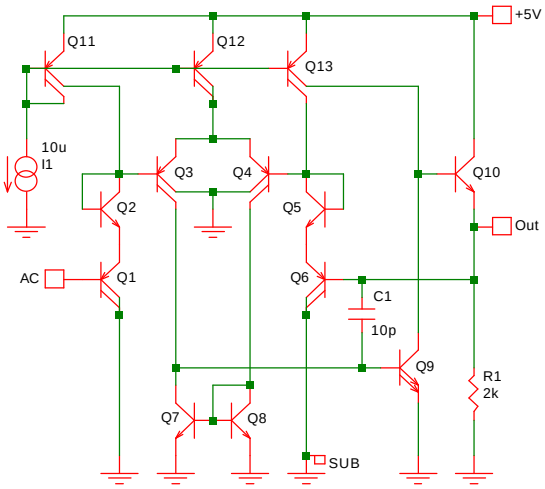


Fig. 16-14: Bipolar half-wave rectifier.

In the circuit of figure 16-14 the output is at ground (without an input signal), held there by R1. If the input moves above ground, the output follows. But if the input goes negative, there is nothing in the output stage that can pull it below ground, so it just stays there.

The value of R1 must be low enough to keep the voltage drop due to the base current of Q6 low. This resistor cannot be replaced with a current sink; the minimum collector-emitter voltage of an NPN transistor is too high.

You can capacitively couple the input signal, with a resistor connected from the input terminal to ground to provide a dc path.

Minimum required supply voltage for a 1-Volt input range is 3.5V.

In figure 16-16 an inverting op-amp configuration with a gain of 1 is used, but it only works for negative-going input signals. As the signal moves above ground, the op-amp is effectively disabled. Thus the output simply follows the input. To avoid loading down the output, a buffer needs to be used. As it happens, the circuit in figure 16-14 is an excellent candidate for this job.

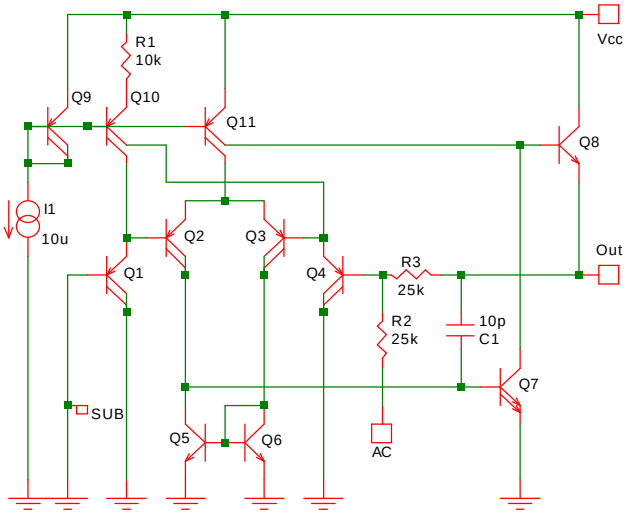


Fig. 16-16: A full-wave rectifier.

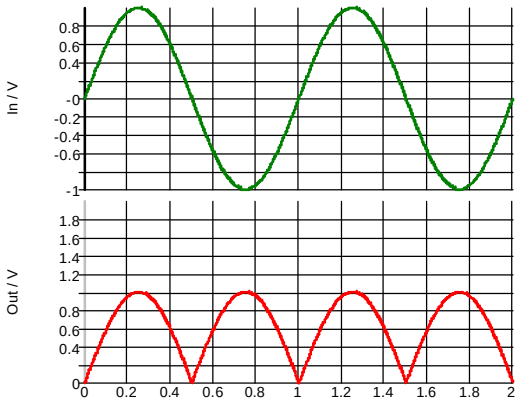


Fig. 16-17: Input and output waveforms for the full-wave rectifier.

Q10 gives a small operating current to Q1 and Q4 (about 1.7uA). Without this, frequency compensation of the op-amp becomes very difficult.

Minimum supply voltage for a 1Vp input is 2 Volts.

Both of these circuits can be readily translated into CMOS. The half-wave rectifier in figure 16-18 uses two advantages of CMOS: first, there is no base current, so the output pull-down impedance can be quite high; second, a current sink can be used at the output instead of a resistor.

With a 1-Volt input range this circuit works down to 1.8V supply at -40°C, 1.6V at 0°C.

The full-wave rectifier of figure 16-19 only needs a 1-Volt supply for the same input range.

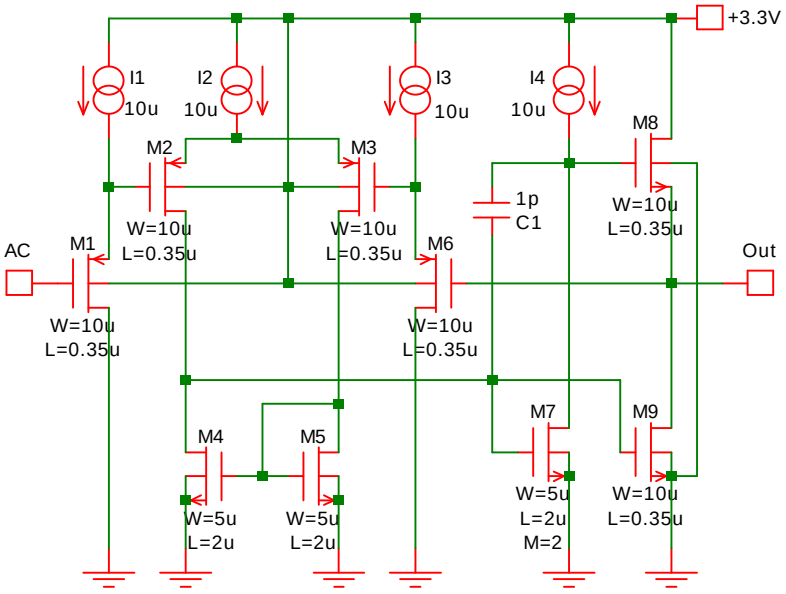


Fig. 16-18: CMOS half-wave rectifier with a single supply.

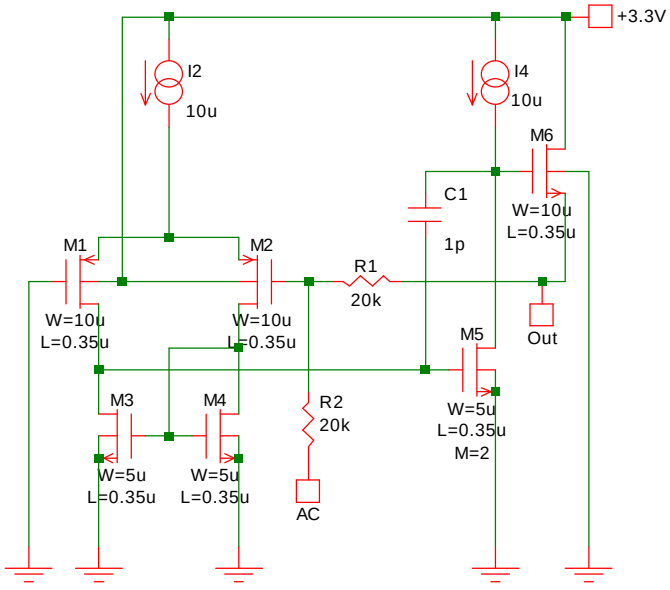


Fig. 16-19: Single-supply CMOS full-wave rectifier.

Averaging the obtained rectified fundamentally takes time; the longer the time constant, the smaller the ripple (but there will always be a ripple).

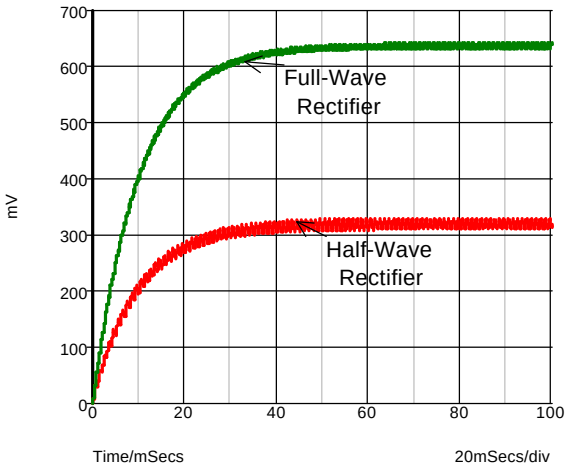


Fig. 16-20: Time constant and ripple for a one-pole low-pass filter used for averaging.

Take the most simple approach, a single low-pass filter (RC) connected to the output. A full-wave output has the advantage, producing less ripple. The time constant used here is 10msec and the signal is 1kHz.

To reduce the ripple you can increase the time constant (which takes longer to reach the final level), or use a higher-order filter.

Thermometers

The PTAT (proportional to absolute temperature) current source has come up before (see page 5-4) . It is unusual and remarkable that we are able to produce a voltage whose value is directly tied to the absolute temperature scale and whose accuracy depends only on ratios, not affected by any process parameter.

Figure 16-21 shows such a circuit. Q1 through Q6 form a loop, started up by leakage alone (see page 5-6). For safety (in case the models are not quite correct) a substantial junction (D1) can be added, which has more leakage current than any of the other devices.

The current in the loop is determined by the emitter ratios of Q2 to Q1 and R1. Recall the formula on page 5-3:

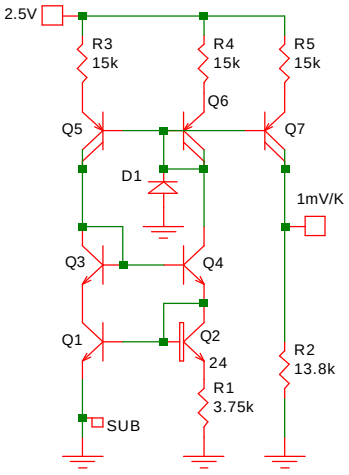


Fig. 16-21: Thermometer with Kelvin scale.

$$deltaVBE = \frac{k*T}{q} * \ln\left(\frac{A1*I2}{A2*I1}\right)$$

I1 and I2 are identical (as produced by Q5, Q6, R3 and R4) and the area ratio is 24. Thus deltaVBE amounts to roughly 83mV at 300K. Since T is in Kelvin, the current increases linearly from zero at absolute zero to 22uA at 300K. This current is then mirrored by Q7 and causes a voltage drop across R2. With the values chosen (and perfect matching), the output voltage amounts to 1mV per Kelvin. Note that any temperature coefficient or absolute variation in R1 is eliminated by a matching R2.

Although the design is relatively insensitive to power supply variation, accuracy is maximized by powering the thermometer from a reference voltage.

Matching is the all-important factor here. A ±1% resistor matching variation will result in an error of ±3°C at room temperature. Adding mismatching of VBE and hFE, you must expect a variation of up to ±5°C untrimmed. With trimming an accuracy of ±0.5°C is possible.

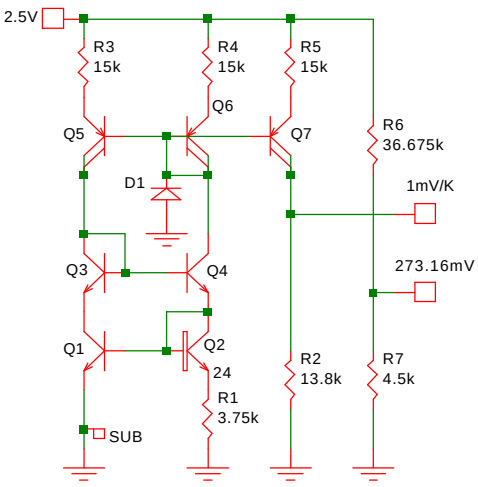


Fig 16-22: A thermometer with a Celsius scale.

The centigrade (or Celsius) scale is the same as the Kelvin one, except for the zero set to 273.16K. So, all we need to do is to create an offset voltage of 273.16mV and read the temperature differentially.

(Similarly we can create a Fahrenheit scale by increasing R2 by a factor of 1.8 and setting the offset to 459.67mV).

Trimming is straightforward: R2 sets the slope; trim it to read 293.16mV at 20°C at the upper terminal; R6 sets 0°C; trim it to read 273.16mV at the lower terminal, also at 20°C.

As we have seen in chapter 7, substrate PNP transistors can be used in CMOS to create a delta-VBE. Figure 16-23 uses this approach to produce a Kelvin output. The current mirror is that of figure 3-25.

The untrimmed accuracy is somewhat worse in CMOS because of the poorer matching of the transistors (for the same area), about ± 7°C at

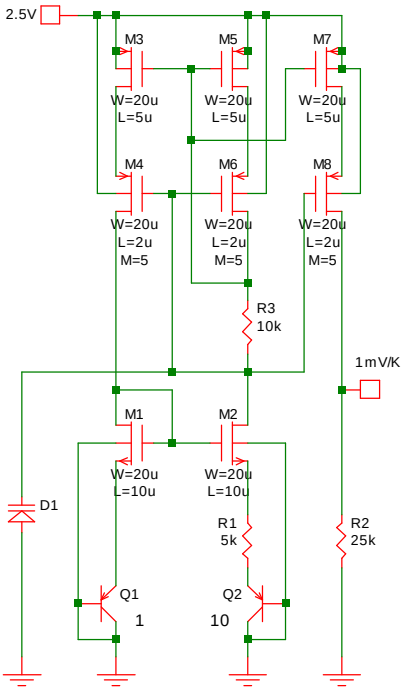


Fig. 16-23: CMOS Thermometer.

room temperature. You can improve this by using a larger emitter ratio for Q2/Q1 and generally larger devices.

But let's not forget the lowly diode. Its forward voltage has a predictable temperature dependence (about -2mV/^{o}C). The slope is subject to absolute variation and not quite as linear as that of a delta-VBE, but the device is nevertheless useful in some applications. For example, if you want to evaluate the temperature at a particular spot on an IC (say next to a power device), use a diode-connected transistor and connect it to a small probing pad. You can then calibrate it first without powering up the chip.

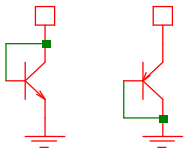


Fig. 16-24: Diode thermometers

Zero-Crossing Detectors

Suppose you need to start a timer or counter at the exact moment when the line voltage crosses the zero line. How do you determine this point without bringing the line voltage into the IC? A simple external resistor will do the trick.

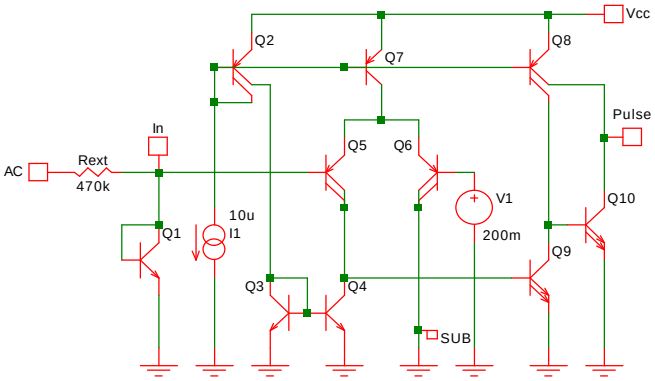


Fig. 16-25: Bipolar zero-crossing detector.

In a bipolar design you can use the (usually unpleasant) fact that a transistor pair used in the first stage can cut off the transistor following in the second stage. In figure 16-25 one input of a differential pair is biased slightly above ground, at about

200mV (derived with a voltage divider from either a reference voltage or from the supply). When the other input is above 200mV, Q9 is turned off and the output is low. As you lower the input voltage, the output goes high at 200mV, when Q5 and Q9 turn on; but as the input drops below ground, Q5 cuts off Q9 and the output drops again. In other words: there is a small window at ground level where the output goes high, otherwise it is always low.

Q1 clamps the positive-going AC voltage so it can do no damage to the IC. The negative-going waveform is automatically clamped by base-substrate diode of Q5. The power dissipation in the external resistor is 25mW.

Be prepared for a surprise when you simulate such a circuit: at first you can't see the output pulse, because it is very small and short compared to the AC waveform.

With 110V and Rext = 470kΩ, the pulse is about 5usec wide. You get the same width at 220V if you double the value of Rext.

The circuit works with a supply as low as 1.2 Volts.

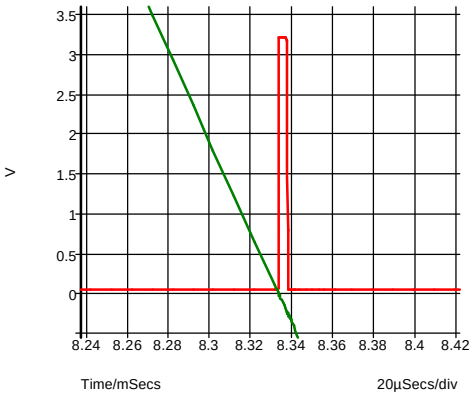


Fig. 16-26: AC slope and output pulse.

The effect used to create the window does not work well in CMOS. Instead we can employ two comparators, one biased at about +200mV (M9, M10) and the other at ground (figure 16-27). Their outputs are then supplied to an "and" gate (M17, M18), which drives the output.

The AC waveform is clamped in the positive direction by two "diode-connected" transistors (M1, M2) and in the negative direction by a substrate diode.

Since there is no base current, you could theoretically make Rext. very large. But also consider that the devices connected to the input (including the pad and the ESD protection device) have a small amount of capacitance. The time constant formed by the external resistor and this capacitance must be smaller than the desired pulse-width.

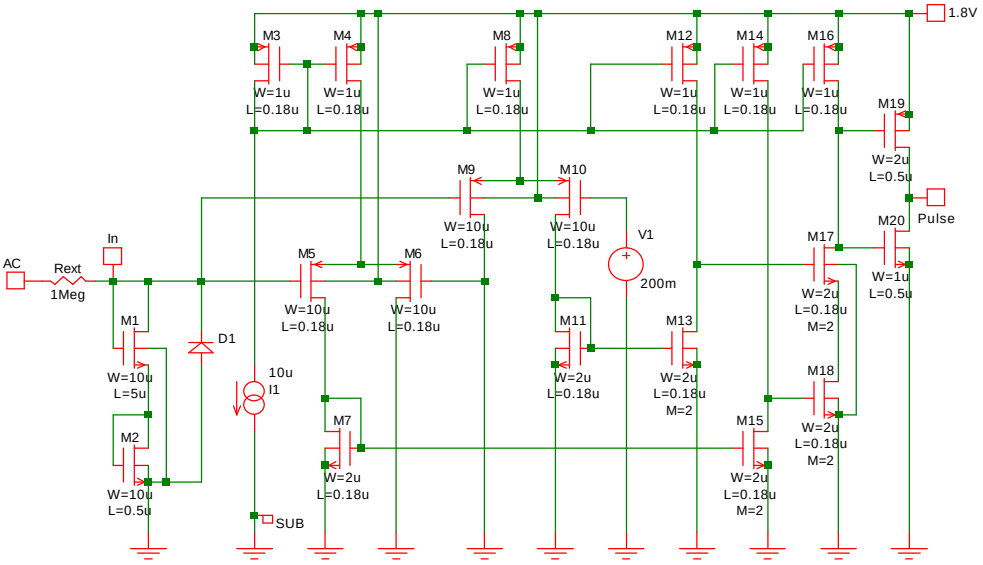


Fig. 16-27: CMOS version of zero-crossing detector.

17 Layout

The layout of analog ICs has so far remained an art, there are no computer programs which could design, place and route the components in an intelligent, competent way. And, more often than not, the person who created the circuit diagram needs to (or should) get involved.

This chapter is by no means a complete guide; it would take an entire book to do the subject justice. Look at it as some hints stemming from practical experience.

Bipolar Transistors

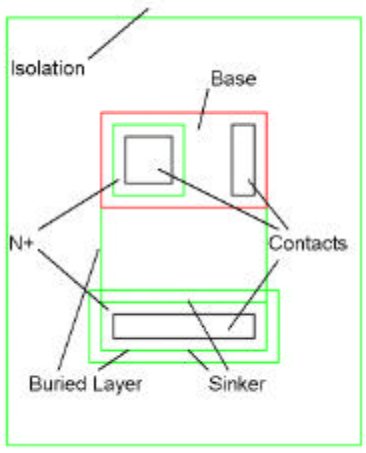


Fig. 17-1: Mask layers for an NPN transistor.

The minimum sequence of masks in a bipolar process is:

Buried Layer
Isolation
Sinker
Base
Emitter (N+)
Contact
Pad

The last mask opens up windows over the bonding pads in a thick glass layer which is spread over the entire circuit to protect the delicate metal.

Note that the emitter (N+) mask is also used to make a low-resistance contact to the collector (the epitaxial layer).

To these seven basic masks several other may be added:

A second (often identical) isolation mask, applied after the buried layer (but before epitaxial growth), to implant p-type regions which diffuse upward (up-down isolation).

A separate mask for high-value (implanted) resistors.

A mask for Schottky diodes, which can consist either of Aluminum or barrier metals directly in contact with the epi layer.

An additional mask for P+ regions, sometimes used to improve the performance of lateral PNP transistors.

A mask for thin-film resistors.

Occasionally a **washed emitter** is used. Here the emitter diffusion (or more likely, implant) takes place through the N+ contact openings (while the windows to the P-regions are masked off). After creating the N+ regions, the thin oxide layer over them is simply etched (or washed) off without a mask. In this way the emitter area can be made smaller, because it is self-aligned with the contact window.

The dimensions of the mask patterns are determined by the process; some of the factors are:

Minimum size of contacts; determined by how small a window can be etched into the oxide. Most of the small-geometry processes require all contacts to be of identical size.

Distance between emitter and base contacts; usually given by the minimum required spacing of the metal covering them.

Overlap of metal over contacts; determined by how well the metal can be aligned to the contact.

Spacing between sinker and base; set by the sideways diffusion of the sinker (and base) and the depletion layer width for the maximum voltage.

Spacing between base and isolation; determined by the sideways diffusion of the isolation (and base) and *two* depletion regions.

Spacing between sinker and isolation; must accommodate two sideways diffusions and one depletion region.

Spacing between buried layer and isolation; must allow for the sideways diffusions of both the isolation and the buried layer. After epitaxial growth the image of the buried layer at the surface is blurred and shifted along the crystal axis (see page 1-17) and thus results in the least accurate alignment.

The isolation mask is a special case. The diffusion takes place *between* the devices, but it would be awkward to draw such a complex web. Thus a convention has been established to draw the isolation region where it is not, and then invert the pattern on the mask.

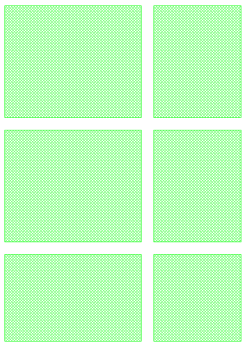


Fig. 17-2: Isolation pattern.

There are several choices for the design of an NPN transistor; figure 17-3 shows some of them.

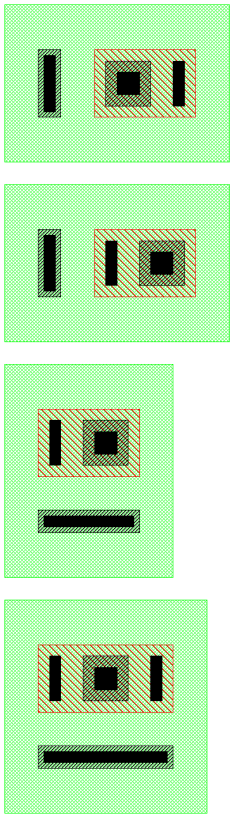


Fig. 17-3: Small NPN transistor patterns.

In the top pattern the emitter is in the center, the base contact on the right and the collector contact on the left. In the pattern below, emitter and base contact are reversed. There is a slight advantage to having the emitter closer to the collector contact, in that the distance the current has to travel in the buried layer to a point underneath the emitter is reduced. (For clarity, the buried layer and sinker patterns have been omitted).

You will also see NPN transistor patterns with more space between emitter and base or base and collector to accommodate metal lines.

In the third pattern the collector contact has been moved, resulting in a somewhat lower saturation voltage.

The bottom pattern contains two base contacts, effectively doubling the current capability of the devices (remember that the maximum current is given by the effective emitter length, i.e. the periphery of the emitter facing the base contact; the rest of the emitter area is ineffective at high currents).

Figure 17-4 shows NPN transistors with two emitters in a single island (or tub). In the top pattern collector and base are common, i.e. connected together, which limits the usefulness of the device.

In the center pattern there are separate bases, only the collector is common. The bottom pattern is identical, with the contacts redrawn for uniform size.

There is a danger with multiple emitters in the same island. As mentioned before, the image of the buried layer is shifted (in all but low-pressure epitaxial processes). The actual buried layer region is where it is supposed to be, a rectangle covering the area from the collector contact (and sinker) to the far edges of the base regions. But the image appearing on the surface is shifted (along the crystal axis). In

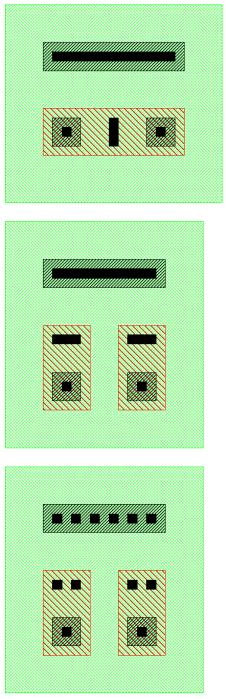


Fig. 17-4: Two-emitter NPN transistors.

<111> silicon starting material, with the wafer-flat at the bottom, the epi-shift is to the right (figure 17-5). The amount of shift is roughly equal to the thickness of the epi-layer. What we see on the surface is a depression, caused by a slight consumption of silicon during the diffusion of the buried layer. Thus it is likely that this step in surface height will hit the left emitter, but not the right one, which influences their matching.

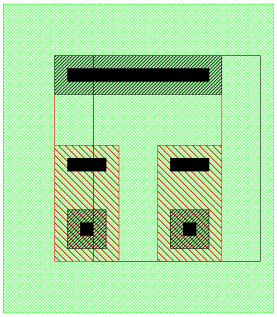


Fig. 17-5: Epi-shift influencing one emitter but not the other.

This effect can be avoided if the entire transistor is rotated so the edge of the shifted pattern falls between the collector contact and the bases. Figure 17-6 shows this with a two-transistor layout. The center transistor has a single emitter, the outer one 16 (used, for example, in a bandgap reference).

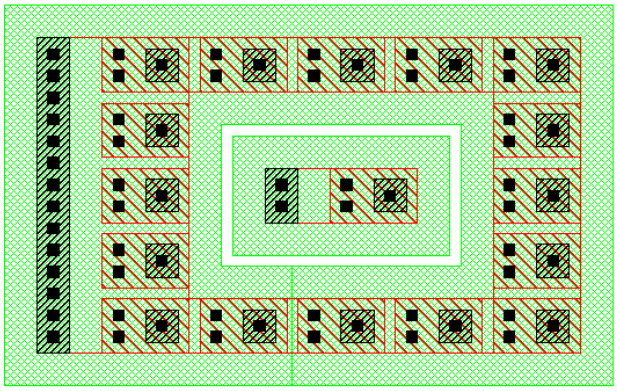


Fig. 17-6: 16:1 emitter ratio with epi-shift mismatch avoided.

The transistor patterns shown so far are all intended for smallest possible size, which naturally limits how much current they can carry. To increase the current capability, the effective emitter length needs to be increased. For the NPN transistor in figure 17-7 not only have the emitters been tripled and stretched, but base contacts have been placed on both long sides. Note that the increase in current capability

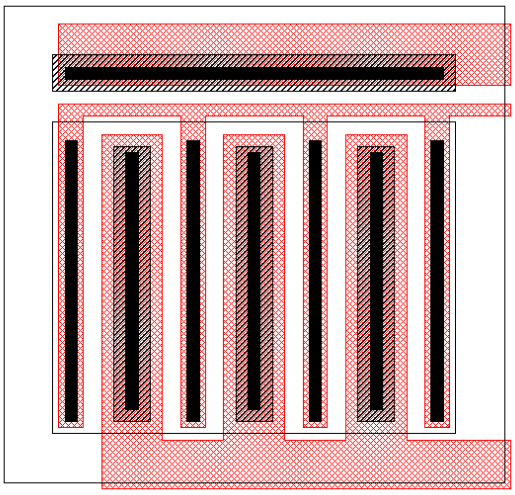


Fig. 17-7: NPN transistor for higher current.

almost always requires wider metal runs for both the emitter and collector (see box on page 14-5).

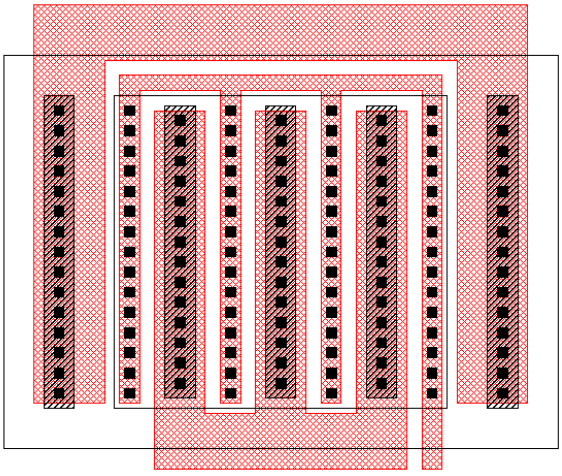


Fig. 17-8: Alternate high-current design.

An alternate design is shown in figure 17-8, with the uniform contact openings required by dense processes. There are two collector contacts (on the outside), three emitters and four base contact columns.

If you increase the size of such a transistor further, there comes a point where it is of advantage to taper both the emitter and collector metal, gradually increasing the width as the currents from more and more contacts are added.

Lateral PNP Transistors

The emitter of a lateral PNP transistor (figure 17-9) is in the center, the dark contact in a p-type (NPN base) diffusion. It is surrounded by the collector, another p-type region. The distance between the outer edge of the emitter and the inner edge of the collector is the base-width. Since both of these regions are on the same mask, emitter and collector are self-aligning and the base-width tends to be very accurate. It needs to be large enough to accommodate the two sideways diffusions and the depletion region spreading from the collector toward the emitter.

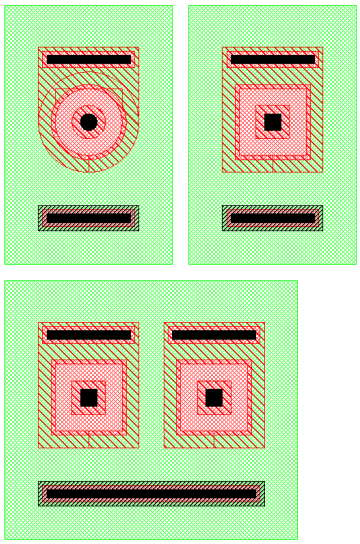


Fig. 17-9: Lateral PNP transistors.

By extending the emitter metal so that it covers the entire base, a field plate is created (always connected to the emitter, which has the highest positive

voltage). This field plate improves the gain of the transistor at low current by keeping p-type charges away from the surface. (In a CMOS process, the poly layer is used as a field plate, also connected to the emitter).

Although a circular emitter results in a uniform base-width and thus (theoretically) produces the highest possible gain, there is actually very little enhancement over the more simple square one.

The third terminal (at the bottom) is the base contact, identical to a collector contact for an NPN transistor.

For a lateral PNP transistor (in a bipolar process) the presence of a buried layer is essential. Without it, the substrate (connected to the most negative supply) would be just as attractive a collector as the intended one; i.e. about half the emitter current would flow to the collector, the other half to the substrate.

The dual pattern at the bottom of figure 17-9 should be avoided. It looks attractive, especially since lateral PNP transistors often have common bases (e.g. in current mirrors), but the two devices influence each other, especially in saturation.

Resistors

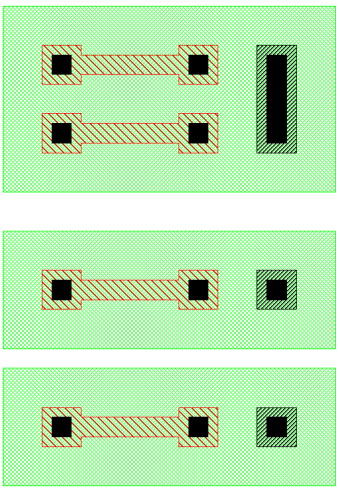


Fig. 17-10: Diffused resistors in a common (top) and in separate (bottom) islands.

In the case of a diffused resistor there is always a surrounding semiconductor region (the epitaxial layer for a bipolar process, a well or the substrate in CMOS). The surrounding region needs to be at a potential so that the junction is reverse-biased. This bias voltage causes a depletion layer to extend into the resistor, decreasing its cross-section and thus increasing its resistance. For a base diffusion this effect is small but occasionally not negligible (about 1%); for an implanted resistor it can be very large (20%).

Thus, if two diffused resistors form a voltage divider, the difference between the bias voltage and the resistor voltage is larger for the lower resistor than the upper one, resulting in a shift of the divider ratio. It may

be small enough to ignore for resistors with 200 Ohms/square (about 0.2%), but for implanted resistors this error is almost always significant. Note that this is an initial error only; it is not subject to change during production.

To avoid this effect, you can place each resistor in its own tub and connect the tub to the positive end of the resistor.

Routing the metal also needs some thought. The Seebeck coefficient (see page 1-31) creates a small voltage between junctions of the same material, located at different temperatures. For this reason it is of advantage to keep connections close together, so that a thermal gradient will have the smallest effect. Figure 17-11 shows a pair of resistors with three sections each, connected on the same side to obtain the shortest distance. For optimum matching in the presence of a thermal gradient the sections also alternate.

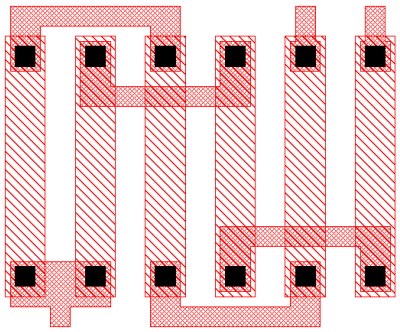


Fig. 17-11: Intermingling and connection of matching resistors.

CMOS Transistors

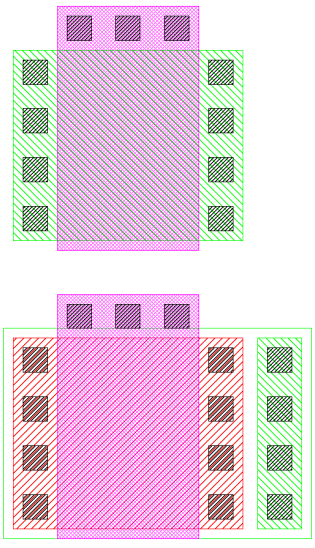


Fig. 17-12: Layout of n-channel (top) and p-channel transistors.

With the basic layers only, the layout of an n-channel transistor is quite simple: there are only three patterns. The first pattern is the poly gate (sitting on top of thin oxide). The second pattern delineates an N+ implant; it is simply a rectangle, protruding on either side of the poly shape. The n-type dopants enter the p-type silicon underneath (the substrate) only outside the gate area; they are stopped by the poly-silicon layer. The third mask places contact opening in the poly and implanted regions.

For a p-channel device two additional masks are required: one for an n-well (surrounding the device or several devices) and one for the P+ implant. The n-well must be contacted and biased.

The patterns in figure 17-12 show long channels, as often required in analog design; the channel length is from left to right, the channel width from top to bottom.

Alas, if things only were that simple. In reality the layout of CMOS IC always involves a large number of masks. There are such layers as field implant, threshold implant, poly 2, poly 3, metal 2, metal 3, metal 4 (an on), interconnections between the metal layers (vias), the pad mask.

Also, some layers are not drawn directly, but are coded. Additional layers are used which form mask patterns only in combination with others.

Most CMOS processes have an n-well (i.e. n-channel devices sit in the common substrate, while p-channel transistors are in common or separate n-wells); sometimes both n and p-wells are present.

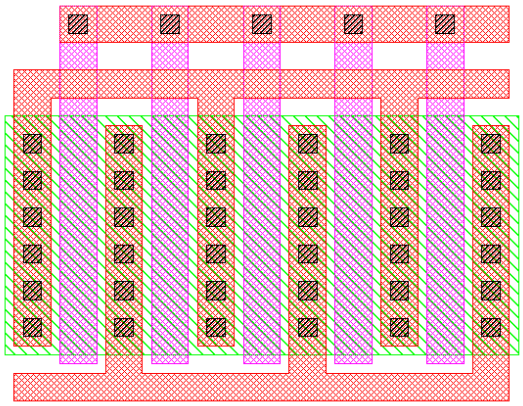


Fig. 17-13: Parallel connection of n-channel transistors.

Drain and source are interchangeable, which leads to a peculiar but efficient way to connect devices in parallel (i.e. to increase the channel width): you connect them in series and merge the terminals between the gates (figure 17-13). Thus the source of one transistor also acts as the drain of the next one, saving space.

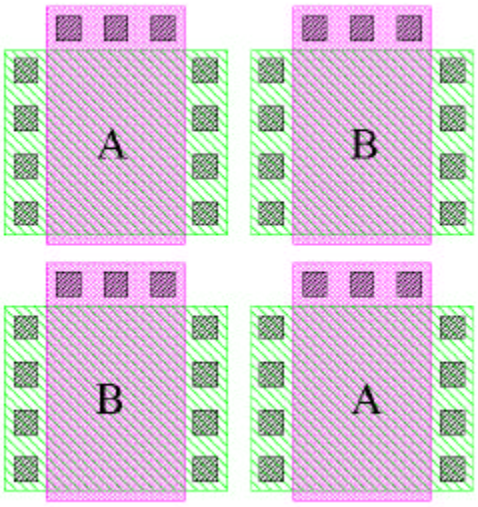


Fig. 17-14: A matching pair of n-channel transistors, using four devices.

The smallest set of matching devices consists of four transistors, arranged to be point-symmetrical. The terminals of the two devices labeled A (figure 17-14) are connected together, as are those for the B transistors. You will find that this is almost impossible without employing the second metal layer.

If there is a gradient (thermal or otherwise) it will affect both devices equally, no matter which direction it takes.

Matching: Myths and Misconceptions

Over the years a number of rules have accumulated around analog design, especially concerning matching devices. For example, most designers believe that matching devices should be intermingled and as close together as possible, because the diffusions or implants have gradients, i.e. vary gradually in depth or concentration over the area of the chip.

A few years ago I had an opportunity to examine this. I measured the matching of adjacent devices and compared that with devices which were farther apart. To my surprise I found no statistically valid difference in matching for a distance of up to 2mm.

It seems that, perhaps, diffusion gradients were present in the early days but, with better furnaces and especially ion implantation, have disappeared to the point where they simply no longer play a role.

To be sure, there are thermal gradients, created by devices which heat one area of a chip more than another. For this reason alone it is wise to intermingle devices and place them close together (and as far from the heat source as possible.

A second belief divides matching devices into as many small pieces as possible, so that they benefit from the statistical effect (large groups of devices match better than two single ones). This has proved to be only marginally true. As you decrease the size of features, the percentage variation becomes greater. Thus, as you approach minimum geometry, matching actually becomes worse for the same overall area.

The third belief holds that you should add dummy devices at the periphery. There appear to be two different explanations to justify this practice: 1. shadows or reflections during exposure act differently on the remote edges than on devices in close proximity, or: 2. the etch-rate for wide spaces is different from narrow ones.

I found no difference between groups of resistors with and without dummy devices at the periphery. It appears that you might be better off using the extra space to make the devices larger.

Cross-Unders

Bipolar analog ICs can often be interconnected using a single metal layer, which lowers the cost. But you will inevitably find spots where two metal lines need to cross.

Using the diffused layer with the lowest resistance (emitter, or N+), one interconnection stops at a contact, dives under the second metal line, and continues at the second contact. This introduces a small amount of resistance (about 20 Ohms), which is tolerable in such places as a base of a transistor.

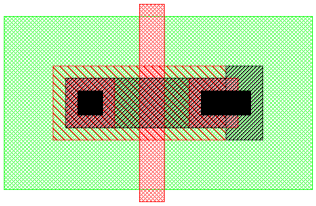


Fig. 17-15: N+ cross-under.

In figure 17-15 the N+ rectangle is placed inside a base diffusion (which sits inside an epi-island). One side of the cross-under is connected to the base region (it does not matter which side, since the voltage drop across the cross-under is bound to be much smaller than that of a diode). If the epi-island is biased at the highest positive supply voltage, you can have several such cross-unders in the same island.

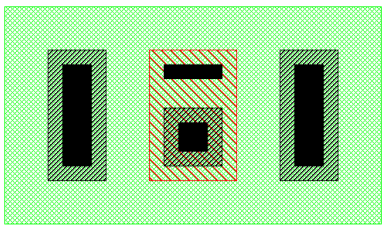


Fig. 17-16: Two collector contacts in an NPN transistors can act as a cross-under.

The epitaxial region can serve as a cross-under as well, though it takes up more space.

A special case in the latter approach is the NPN transistor with two collector contacts (figure 7-16). Here a line connected to the collector stops at the right-hand contact and continues on at the left-hand one.

But be careful with this scheme. Let's assume the resistance between the contacts is 100 Ohms. The resistance between one contact and the center (i.e. a point underneath the emitter) is then about 50 Ohms. If one contact carries all or most of the collector current, the second contact will display the voltage at the center and not that of the first contact.

Kelvin Connections

Any contact on the surface of an IC has some resistance, which often makes precision measurements difficult. This can be avoided by providing two sets of contacts, one to carry the current, the other to measure the voltage.

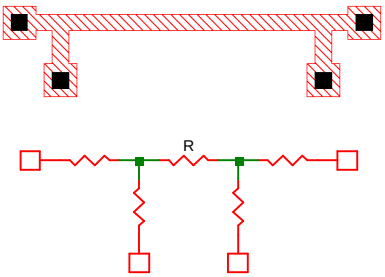


Fig. 17-17: A Kelvin connection for a resistor and its equivalent circuit.

In a Kelvin connection contact resistance is of no consequence. The two resistances in the current path add a bit to the headroom required, the two in measurement path simply need to be negligible compared to the measuring impedance.

Metal Runs and Ground Connections

The concept of an "analog ground" is often misunderstood. It is meant to be a noise-free point (or hub), a spot either on the circuit board or on the IC which can be used as a 0-Volt reference.

The usual practice designates a pin which carries little or no current as the analog ground; other pins, intended to be at the same potential but carrying current are then connected to this point on the circuit board.

There is another way to achieve this, one which saves a pin and has better performance. A package pin has low resistance, lower than a trace on a circuit board or a metal run on the IC. Designate a pin as the analog ground and then connect not one but two neighboring pads to it with separate bonding wires: one carries no current and serves as the analog reference ground on the IC, the other carries the potentially polluting currents.

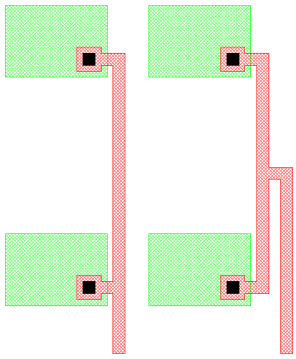


Fig. 17-18: Proper connection (on the right) for matching devices.

Similarly, on the IC, use separate metal runs to connect sensitive devices. In figure 17-18 the left-hand connection can create an error. Assume the runs lead to emitters carrying 1mA. With say 50 squares (at 30mOhms per square) of additional aluminum for the upper device, the voltage drop is 1.5mV, creating a current mismatch of 6% at room temperature. With the balanced connection on the right this is avoided.

Back-Lapping and Gold-Plating

To fit into small, shallow packages, wafers are often thinned down by back-lapping (a somewhat messy, wet grinding operation). This removes not only the oxide layer on the back but any diffusions which may have taken place there, giving direct access to the substrate material. If you add a gold-plating step you get a low-resistance connection directly to the substrate.

Ordinarily such a connection is not essential (the substrate is also contacted from the top). But if you have sinned and allowed high substrate currents, you may be able to suppress the resulting effects in this way.

DRC and LVS

A computer is clearly not as smart as a human being (or so we like to think), but it is adamantly intolerant of errors and it never tires.

There are two checking operations required when a layout is finished. The first is Design Rule Checking (DRC), where it is made sure that the dimensions and spacings in each layer in each device and for all the connections obey the design rules.

The second compares the layout with the (simulated) schematic (layout versus schematic, or LVS).

One has to face the humiliating fact that the human being is not well suited for either job. Great attention to an excruciating amount of detail is required, which we cannot handle without making mistakes.

Only if DRC and LVS are done by computer can you be certain that the chip will contain what you think it contains.

* * *

Thus ends this book of the minority field in the world of semiconductors. A field past glamour, often neglected, but undeniably essential. And a field of great satisfaction for those who know it.

References

Chapter 1

History of the Transistor:

Pearson, G.L. and Brattain, W.H.: "History of Semiconductor Research", Proceedings of the IRE, December 1955, pp. 1794-1806

Shockley, W.: "The Invention of the Transistor", National Bureau of Standards Publication # 388, May 1974

Shockley, William: "The Path to the Conception of the Junction Transistor", IEEE Transactions on Electron Devices, July 1976, pp. 597-620

Brattain, Walter H.: "Genesis of the Transistor", The Physics Teacher, March 1968, pp. 109-114

History of the Integrated Circuit:

Wolff, Michael F.: "The Genesis of the Integrated Circuit", IEEE Spectrum, August 1976, pp. 45-53

Interviews with Phil Ferguson, Victor Grinich, Jean Hoerni, Eugene Kleiner and Robert Noyce, 1983

Reid, T.R.: "The Chip", Simon and Schuster, 1984

Transistor Design:

Muller, Richard and Kamins, Theodore: "Device Electronics for Integrated Circuits", John Wiley and Sons, 1977

Roulston, David: "Bipolar Semiconductor Devices", McGraw-Hill, 1990

Chapter 2

Antognetti, Paolo and Massobrio, Guiseppe: "Semiconductor Device Modeling with Spice", McGraw-Hill, 1988

Kundert, Kenneth: "The Designer's Guide to Spice and Spectre", Kluwer Academic Publishers, 1995

Berkeley BSIM model information is available at www-device.eecs.berkeley.edu/~bsim3/

Chapter 5

Fig. 5-7: This circuit is usually attributed to Bob Widlar, but in his paper and patent he considered only 1:1 emitter ratios. Widlar, "Some Circuit Design Techniques for Linear Integrated Circuits," IEEE Transactions on Circuit Theory, Dec. 1965, pp. 586-590. Widlar, "Low-value current source for integrated circuits," US Patent 3,320,439, 1967

Fig. 5-14: George Erdi, "Starting to Like Electronics in Your Twenties", p 172, in Williams, "Analog Circuit Design", Butterworth-Heinemann, Stoneham, MA, 1991. Erdi, US Patent 4,837,496, 1989

Chapter 6

dB: Martin, W.H., "DeciBel - The New Name for the Transmission Unit, Bell System Technical Journal, January 1929

Steinmetz: Wagoner, C.D., "Steinmetz Revisited", IEEE Spectrum, April 1965, pp.82-95. Kline, Ronald R., "Steinmetz", Johns Hopkins University Press, 1992

Fourier: Encyclopedia Britannica

Chapter 7

Hilbiber, D.F., "A New Semiconductor Voltage Standard", International Solid State Circuits Conference, 1964 (ISSCC 1993 Commemorative Supplement, pp. 34-35)

Widlar, R. J., "New Developments in IC Voltage Regulators", ISSCC Digest of Technical Papers, Feb. 1970, pp.32-33, and IEEE Journal of Solid-State Circuits, Feb. 1971, pp. 2-7

Brokaw, A. P., "A Simple Three-Terminal IC Bandgap Reference", IEEE Journal of Solid-State Circuits, December 1974, pp. 388-393. Brokaw, "Solid-State Regulated Voltage Supply", US Patent 3,887,863, June 3, 1975

Widlar, R.J., "Temperature Compensated Bandgap IC Voltage References", U.S. Patent 4,249,122, Feb. 3, 1981

Gunawan, M., Meijer, G., Fonderie, J. and Huijsing, J., "A Curvature-Corrected Low-Voltage Bandgap Reference", IEEE Journal of Solid State Circuits, June 1993, pp. 667-670

Chapter 8

Solomon, James E.: "The Monolithic Op Amp: A Tutorial Study", IEEE Journal of Solid State Circuits, December 1974, pp. 314-332

Figures 8-10, 8-11 and 8-14: Output stage and base current compensation scheme were derived from Linear Technology Corporation's LT6011

Hogervorst, Ron, et al: "A Compact Power-Efficient 3V CMOS Rail-to-Rail Input/Output Operational Amplifier for VLSI Cell Libraries, IEEE Journal of Solid State Circuits, December 1994, pp. 1505-1513

De Langen, Klaas-Jan and Huijsing, Johan H.: "Compact Low-Voltage Power Efficient Operational Amplifier Cells for VLSI, IEEE Journal of Solid State Circuits, October 1998, pp. 1482-1496

Chapter 10

Figure 10-2 is derived from the National Semiconductor LM13700. A similar concept is used in the RCA (now Intersil) CA3280

Figure 10-7: Nedungadi, A. and Viswanathan, T.: "Design of Linear CMOS Transconductance Elements", IEEE Transactions on Circuits and Systems, October 1984, pp. 891-894

Chapter 11

Low-Voltage 555: Camenzind, Hans R.: "Redesigning the old 555", IEEE Spectrum, September 1997, pp. 80-85

Matthys, Robert J.: "Crystal Oscillators", revised edition, Krieger Publishing Company, 1992

Chapter 12

Gardner, Floyd M.: "Phaselock Techniques", John Wiley and Sons, second edition, 1979

Grebene, Alan B. and Camenzind, Hans R.: "Frequency-Selective Integrated Circuits Using Phase-Lock Techniques", IEEE Journal of Solid State Circuits, August 1969, pp. 216-225

Chapter 13

Sallen, R.P. and Key, E.L.: "A Practical Method of Designing RC Active Filters", IRE Transactions on Circuit Theory, March 1955, pp. 74-85

Butterworth, S.: "On the Theory of Filter Amplifiers", Wireless Engineer, 1930, pp. 536-541

Brodersen, R.W., Gray, P.R. and Hodges, D.A.: "MOS Switched Capacitor Filters", Proceedings of the IEEE, January 1979, pp. 61-75

Chapter 14

Pressman, Abraham I.: "Switching Power Supplies", McGraw-Hill, 1991

Billings, Keith: "Switchmode Power Supply Handbook", McGraw-Hill, 1989, 1999

Camenzind, H.R.: "Modulated Pulse Power Amplifiers for Integrated Circuits", IEEE Transactions on Audio and Electroacoustics, September 1966, pp. 136-140

Attwood, Brian E.: "Design Parameters Important for the Optimization of Very-High Fidelity PWM (Class D) Audio Amplifiers", Journal of the Audio Engineering Society, November 1983, pp. 842-853

Duncan, Ben: "High Performance Audio Amplifiers", Newnes, 1996

Chapter 15

Analog Devices: "Analog-Digital Conversion Handbook", Prentice-Hall, 1986

Van de Plasche, Rudy: "Integrated Analog-to-Digital and Digital-to-Analog Converters", Kluwer, 1994

Boser, Bernhard E. and Wooley, Bruce A.: "The Design of Sigma-Delta Modulation Analog-to-Digital Converters", IEEE Journal of Solid-State Circuits, December 1988, pp. 1298-1308

Candy, James C. and Temes, Gabor C.: "Oversampling Delta-Sigma Data Converters", IEEE Press, 1992

Analog Devices: "Sigma-Delta ADCs and DACs", Application Note AN-283

Chapter 16

Gilbert, Barrie: "A New Wide-Band Amplifier Technique", IEEE Journal of Solid-State Circuits, December 1968, pp. 353-365

Gilbert, Barrie: "A High-Performance Monolithic Multiplier Using Active Feedback", IEEE Journal of Solid-State Circuits, December 1974, pp. 364-373

Index