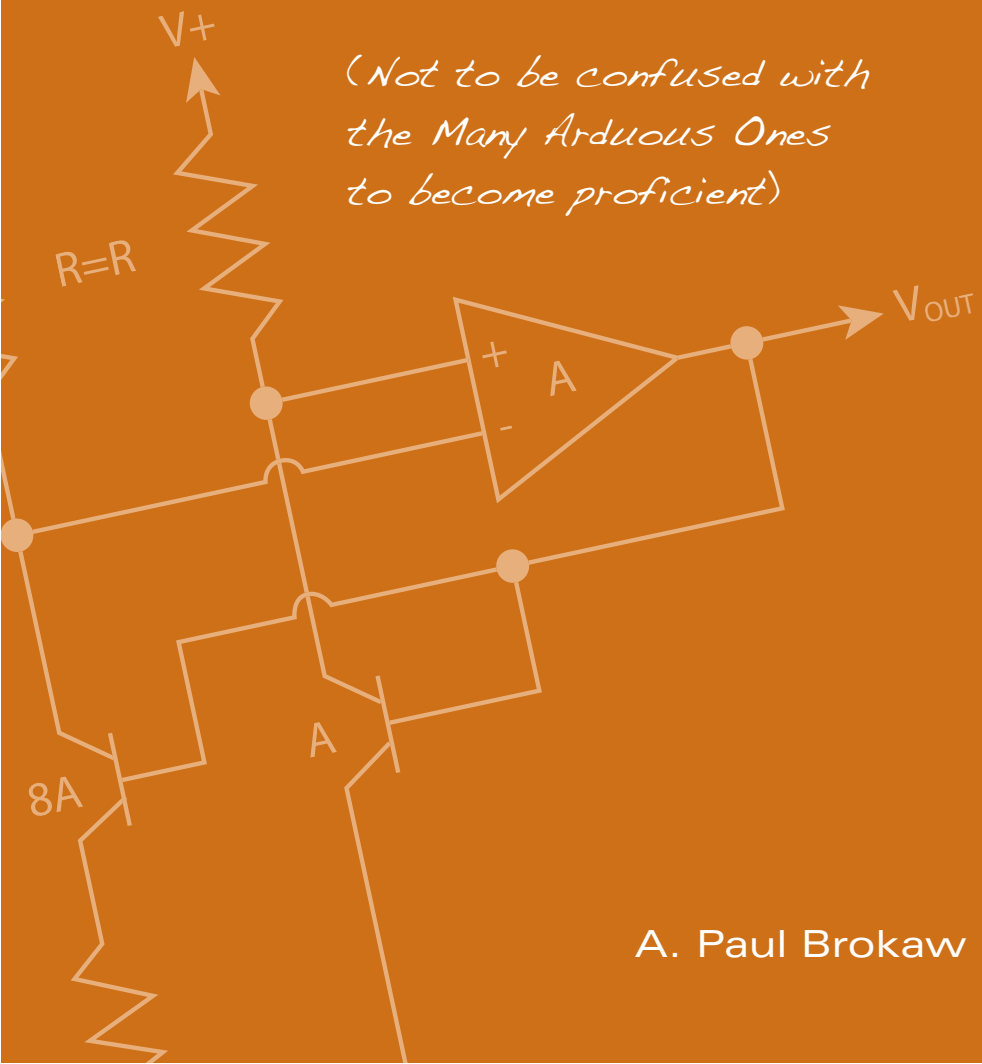
How to Make a

BANDGAP VOLTAGE REFERENCE

in ONE EASY Lesson

(Not to be confused with the Many Arduous Ones to become proficient)



A. Paul Brokaw

A voltage reference is not only a convenience, but is also a necessity for many electronic devices, for example a voltage DAC which converts a digital input to an output voltage. This result is the product of the digital word, which is scalar with no dimension, and some voltage to which the scaled output is referred. The DAC needs to "know" how big a certain voltage is, in order to set its full scale value.

We need a component which "knows" how big a volt is.

It turns out that the silicon all around us can be persuaded to give up its secret knowledge of the Volt if we will accommodate it.

The base emitter voltage of a bipolar transistor versus temperature can, in theory, be extrapolated to equal a known physical constant which has the dimension Volt at a temperature of zero Kelvin.

This constant is called the extrapolated bandgap of silicon.

Unfortunately, this voltage is temperature sensitive, but predictably so.

However, silicon has a second property which also relates temperature and voltage and this can be combined with the change in Vbe to almost cancel the temperature effects and make a voltage approximating the bandgap at all temperatures, or at least the temperatures most of us require.

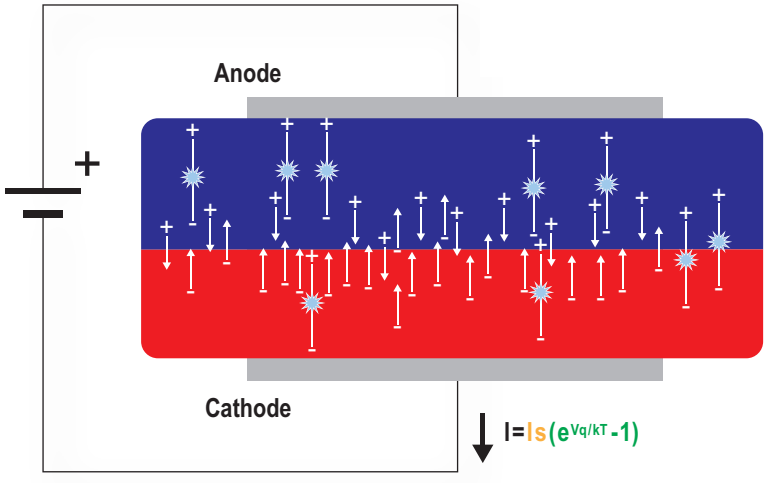
LET ME BEGIN with a sort of hand-waving description of PN junctions and bipolar transistors. This representation of a PN junction is shown forward biased. Positive voltage on the P-type anode (blue,) drives holes toward the junction with N-type (red) material. At the same time the negative voltage on the N-type material repels electrons toward the junction where many recombine with the holes. Some of the electrons may cross over into the P-type, where they are minority carriers that do not spontaneously arise in P-type silicon. Similarly, holes cross the junction and become minority carriers in the N-type. Minority carriers will recombine with carriers of opposite polarity, within the silicon.

The carriers that enter the silicon and combine with opposite polarity carriers do not re-emerge at their injection terminal, and so constitute a net current flow.

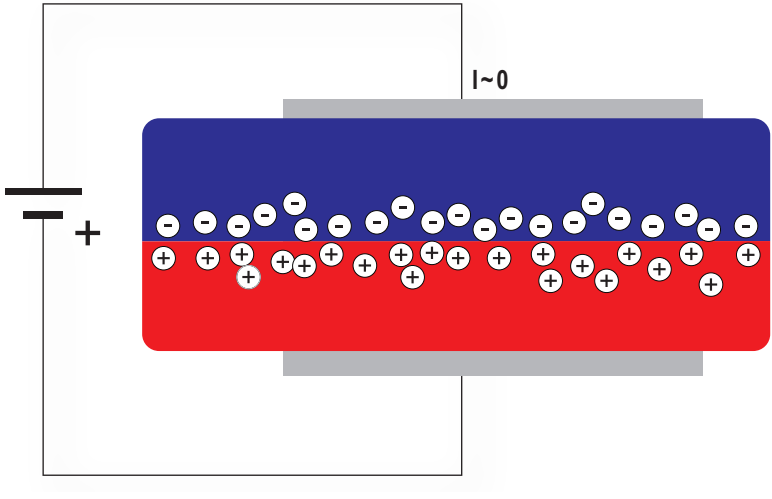
The quantities of the two minority carrier types aren't necessarily equal, and the ratio is related to the amount of doping of neutral silicon with the respective dopants which make the regions P and N.

The carrier injection will depend upon the voltage applied to the diode, and so the current flow will be voltage dependent, though not linearly so, as you should expect of a resistor.

If we reverse the polarity of the applied voltage the current flow will be greatly reduced to near zero. This results from the carriers being pulled away from the junction, leaving the dopant atoms, which are fixed in the silicon lattice. These fixed dopants aren't free to move and recombine. The small current that flows is due only to thermal ionization, and we will neglect it.



Forward Biased P-N junction



Reversed Biased P-N junction

Now, suppose we restore the forward bias to the junction, causing a current to flow. Next, let's add another N-type region within the P-type, and drive it in until it is separated from the other N region by a thin layer of P-type material and apply to it a bias voltage greater than the voltage on the forward biased junction.

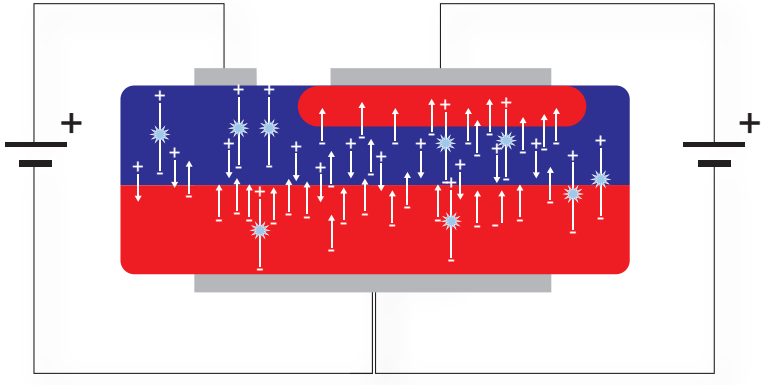
Since the added N region is more positive than the P region, the new junction is reverse biased, and we might expect very little current to flow. However, some of the electrons pulled into the P region will be attracted by the added N region, and they will cross that junction, become majority carriers in the N region, and will constitute a current flow.

If we take as our goal to maximize the electron flow from bottom to top, there are several things we can do. One is to reduce the size of the bottom N region so that all the minority carriers will be less likely to recombine in the P region. Another is to adjust the doping levels to favor more electrons and fewer holes crossing the lower junction.

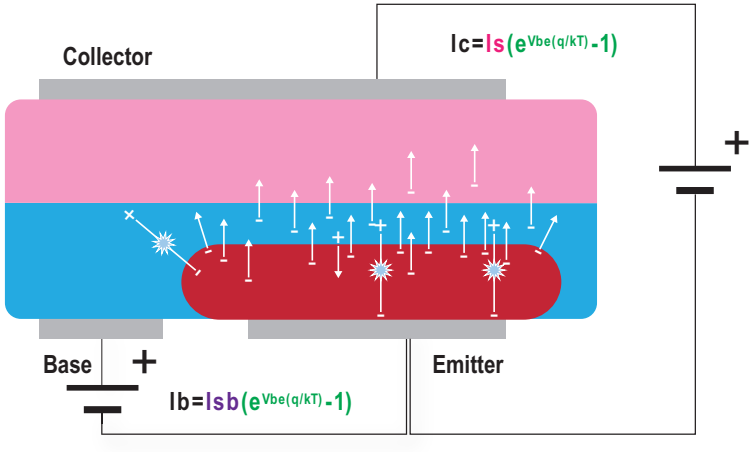
By now I'm sure you recognize the Emitter, Base and Collector of an NPN and we have maximized electron flow from emitter to collector by making the P-type base thinner and covering the emitter with the collector junction, and have reduced the base current by doping to reduce the number of holes injected into the emitter.

This seems a good place to interject a personal preference and the reasons for it. First, please note that the story I just told shows the base voltage to be the cause of both base and collector currents. The geometry and doping is selected to minimize the base current as a fraction of emitter current, or to raise beta, the ratio of collector current to base current.

Because the transistor construction will control beta, we are free to consider either base voltage or base current as the cause of collector current.



Additional diffusion forms Crude NPN Transistor



Adjustments to size and doping makes improved NPN transistor

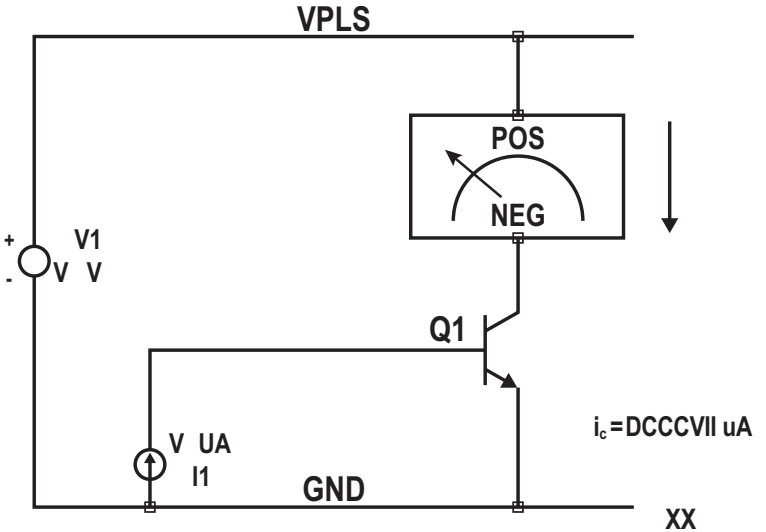
However, I will assert that most of the time (with exceptions granted) it will be easier to treat base voltage as the cause of collector current.

Here's an example of the principle involved. Suppose we measure the base and collector current of a transistor and record our measurements in Roman numerals. Then to determine beta we need only divide one by the other to discover that beta = CLXI+IV/X.

Now, that result is perfectly valid and as accurate as our measurements, but wouldn't it have been so much easier if we had used our modern number system?

Well, that's similar to doing an analysis using beta vs. Vbe. You can get the right answer either way, but as we'll see one is much easier than the other.

An additional issue is that in most applications you want to minimize sensitivity to beta or base current. Why would you want to minimize the effects of your major controlling transistor parameter?



beta = $\frac{\text{DCCCVII}}{\text{V}}$;

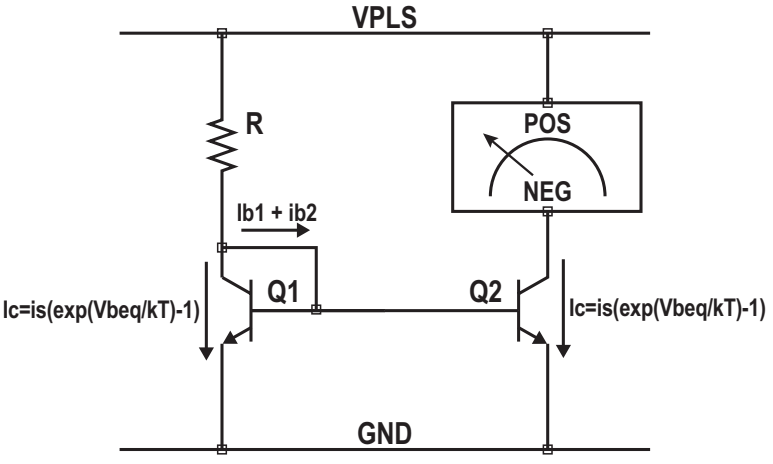
```
                XX
                XX
                XX
              C     I
V ) DCCCVII
  D
    C
    C
      C
      C
        C
        C
          V
          V
            II
```

XX+XX+XX = LX & II/V = IV/X ⇒ beta = CLXI+IV/X

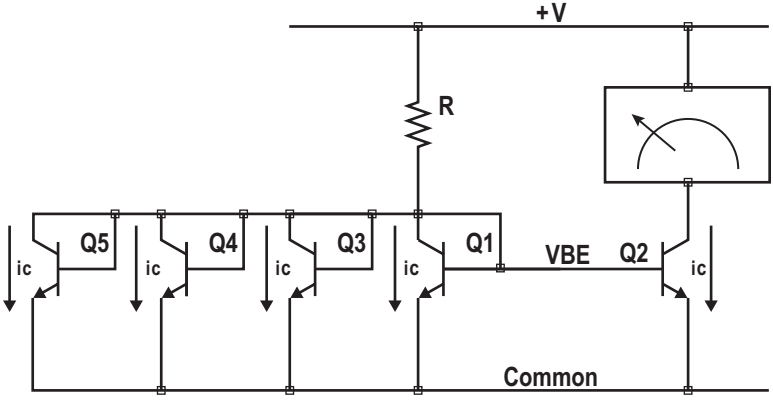
http://www.guernsey.net/~sgibbs/roman.html

With that observation in mind, let's look at a simple current mirror. Q1 and Q2 have the same Vbe, so we can hope they will have the same collector current. The collector current of Q2 may be higher than Q1 since the collector voltage is higher. The collector current of Q2 may also be reduced from the current in R because the two base currents subtract from R current before it drives the collector of Q1. In this case we would treat the base currents as an error in current mirroring along with the error due to the mismatch of collector voltage. These errors have opposite signs, but let's see if we can't do better.

Another simple circuit will reduce by 4X the input current mirrored by Q2, with the same approximations. Note that all the transistors are alike, with the same Vbe so that they all operate at the same current density or current per unit of active emitter area.



Simple Current Mirror



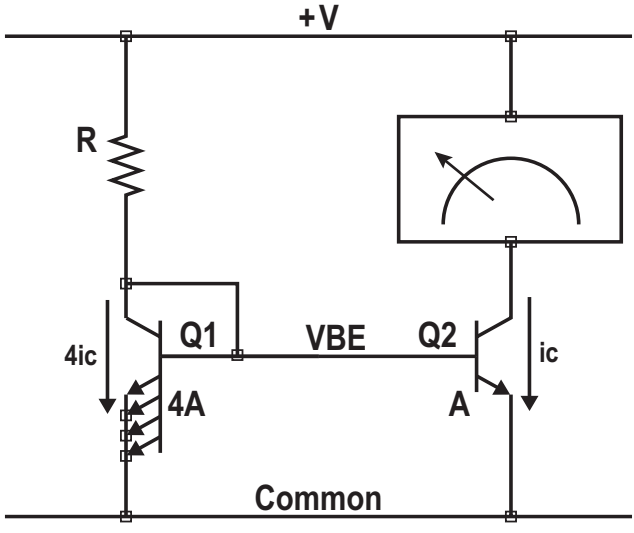
A 4:1 mirror operates all transistors at the same current density

A sort of shorthand we might substitute for the explicit four transistors is to show a single transistor collector and base with four emitters. The significance of this symbol is shown in the next picture, which is a cross section of such a transistor.

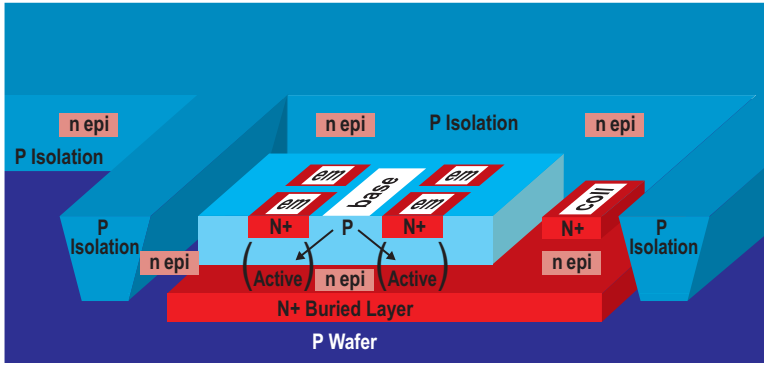
This cross section shows a P-type wafer with N-type epi and a device including an N-type well with buried layer and a P ring to isolate this region from the rest of the epi. A P-type base is diffused into the isolated well and heavily doped N+ dots are diffused into the base. At the left of the base the N+ dot, the P-type base, and the underlying n-epi form an NPN transistor. This transistor is confined to the region below the emitter and shares isolation and base with three others

There are four emitter dots diffused into the base and so this device acts like four NPNs the base and collectors of which are joined by construction, and with four separate emitters.

The emitters are separated, rather than combined into one large one, so they may be well matched to a device with a single identically sized emitter. Typically for bandgap circuit applications the four emitters will be joined by metallization, but in other applications they may be used individually.



Pictorial shorthand represents construction

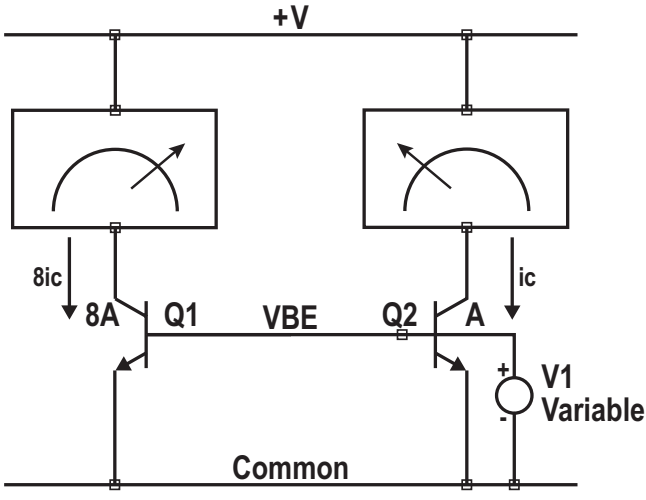


Junction Isolated NPN with multiple emitter sites

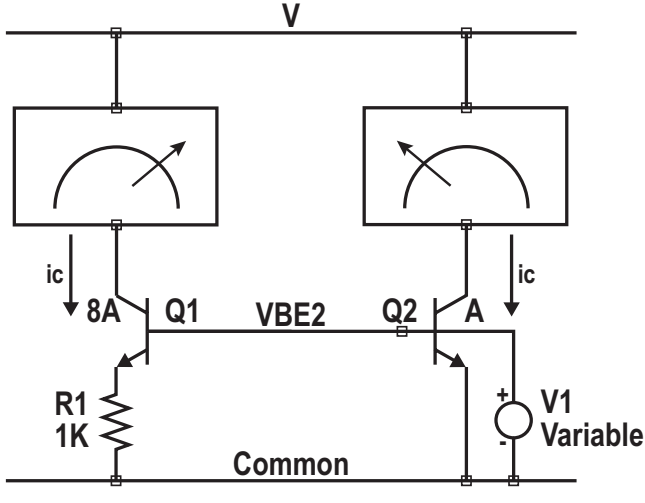
Even this simplification of the symbol becomes clumsy when more emitters are wanted and so a multi emitter device may be indicated by notation nA and to show what each emitter matches, one is designated A.

In this diagram both bases are driven by a common voltage V1, so the circuit is no longer a current mirror, but is a source of two collector currents, instead. If we trust that the currents are a function of base voltage without evaluating the direct relation, we can note that some voltage V1 applied to Vbe 1 should result in a particular collector current from Q2. Since Q1 is effectively eight (I've chosen n=8, somewhat arbitrarily) transistors in parallel, all of which have the same Vbe as Q2, I expect the net collector current of Q1 to be eight times that of Q2. Since they share collector, base and emitter voltage, the systematic error in this ratio should be very small, and in particular, unaffected by base current, which is supplied to both transistors by V1.

Now let's introduce a resistor (1K) into the emitter connection of Q1. As the current is adjusted by changing V1 the corresponding voltage drop across R1 will disturb the ratio of collector currents.

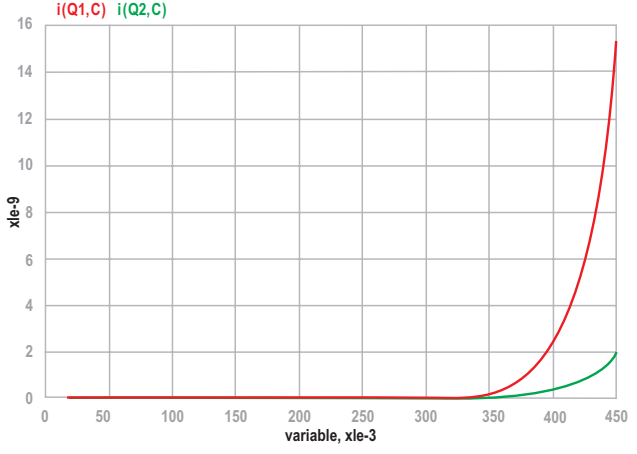


A further condensation uses 8A and A to indicate an 8:1 ratio of the two outputs of this voltage-driven current source

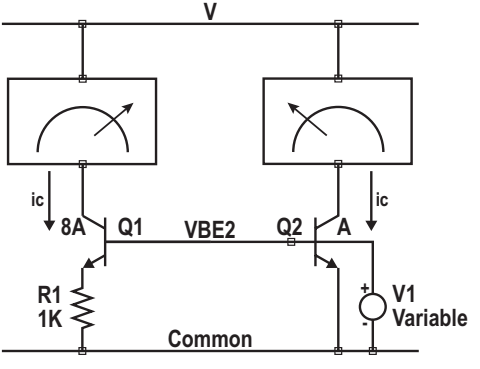


An added resistor changes the emitter voltage of Q1

For small values of V1 where currents are relatively small, the voltage across R1 will be small and will hardly interfere with Vbe of Q1, so the collector currents will stay very nearly in the ratio eight to one. I hope you will take my word for it below 350mV, and note that at 450mV the eight to one ratio is only starting to be reduced by the difference in the two Vbes resulting from ~15 nA in R1.



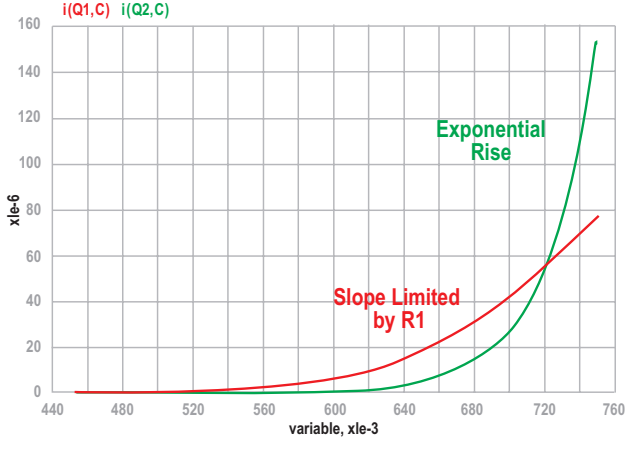
When the resistor voltage is small, the collector current ratio is close to 8:1



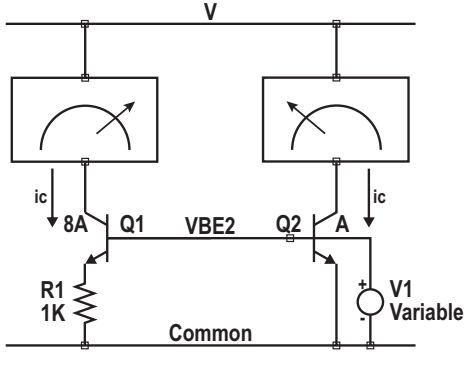
But, as we will see, while the current in Q2 can rise exponentially with V1, the rise of current in Q1 will be asymptotic to a linear slope defined by R1.

This means that as V1 rises, the ratio of collector currents falls until it is below one. In the figure this happens as V1 crosses ~720mV in a room temperature simulation.

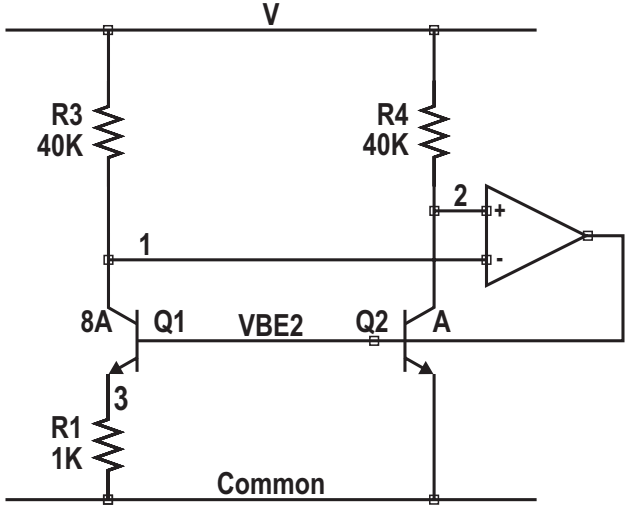
In order to more accurately set and maintain this crossover voltage an opamp may be used to adjust Vbe2 to the value that makes the collector currents equal.



At higher currents Vbe1<Vb2 and the current ratio falls



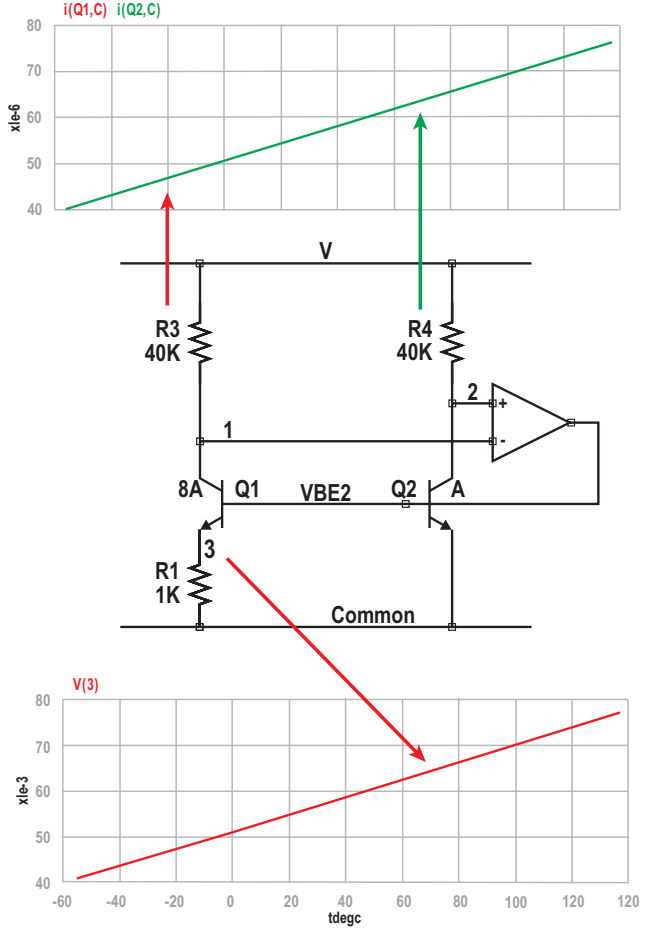
An op-amp drives the common base line to maintain Equal collector currents



In this arrangement both collector currents are driven up and down together, but their different slopes result in a differential signal at the opamp input. There is a negative feedback path including Q2, node 2, and the opamp, as well as a positive feedback path including Q1 with R1, node 1, and the opamp. R1 reduces the effective Gm in the positive feedback loop which reduces the gain of that path below that of the negative feedback path. So, when the two currents match and nodes 1 and 2 are equal in voltage, the negative feedback dominates and Vbe2 is maintained at the voltage which keeps the collector currents balanced. It will be important to note that by making the collector currents equal we have made the current densities unequal by a factor n = 8.

While the feedback makes the two collector currents equal it does not answer the question of what value the currents actually have. However, it should be apparent that the common value for the currents is closely approximated by the current in R1, which itself is determined by the difference between Q1 Vbe and Q2 Vbe.

As indicated before, the temperature of the devices affects the result. This figure shows that V3, the voltage across R1, changes in proportion to temperature. And the resulting current appears at the two collectors. The currents track so well that the Q1 current line on the graph is obscured everywhere by the Q2 current trace.



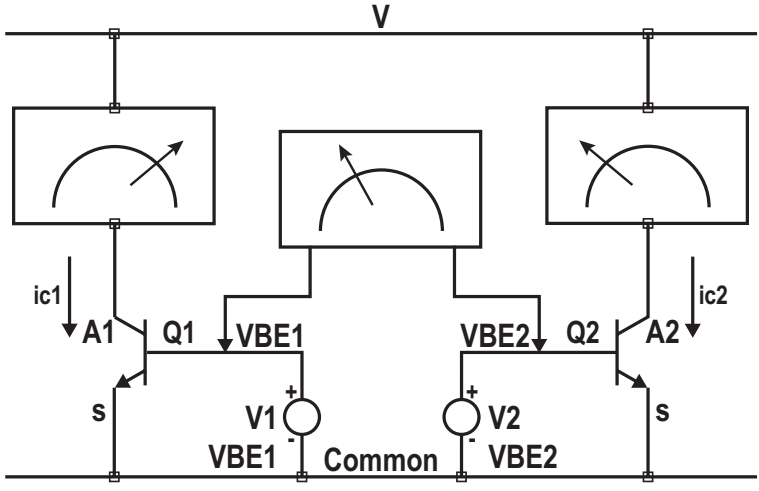
The two collector currents track almost perfectly as temperature is swept – And their common value increases with rising temperature as indicated by the voltage across R1

To see what and why the voltage across R1 is what it is, let's examine a little experiment. Take two identical NPN transistors and bias each into conduction with a voltage applied to the base, using a different voltage for each device.

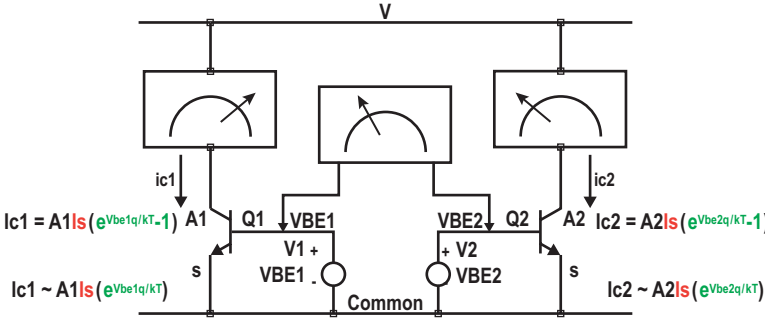
Arrange to measure each collector current and the *difference* of the base voltages.

Now, using the junction equation for the two transistors, the form is the same, but allowance is made for the ratios of current and area. One of the factors in the collector current is an exponential function of Vbe and q, electron charge, over k, Boltzmann's constant, and T, Absolute Temperature, all minus one. Since the exponential will be on the order of 1e14 to 1e18 for the transistors and environment we will use, the subtraction of one can safely be neglected.

Take the ratio of current densities and with a bit of manipulation you will find the difference of the Vbes or ΔVbe as it is generally referred to, depends upon the current ratio which is controlled in the circuit, the emitter area ratio, Kelvin temperature and the constants.



Current Density vs. Base Voltage Difference



Then: $(ic1/A1)/(ic2/A2)=(e^{Vbe1q/kT})/(e^{Vbe2q/kT})=e^{(Vbe1-Vbe2)q/KT}$

So: $(q/KT)(Vbe1-Vbe2)=\ln((ic1/A1)/(ic2/A2))$

$\Delta Vbe=Vbe1-Vbe2=(kT/q)\ln((ic1/A1)/(ic2/A2))=(kT/q)\ln((ic1/ic2)(A2/A1))$

Current Ratio — Area Ratio

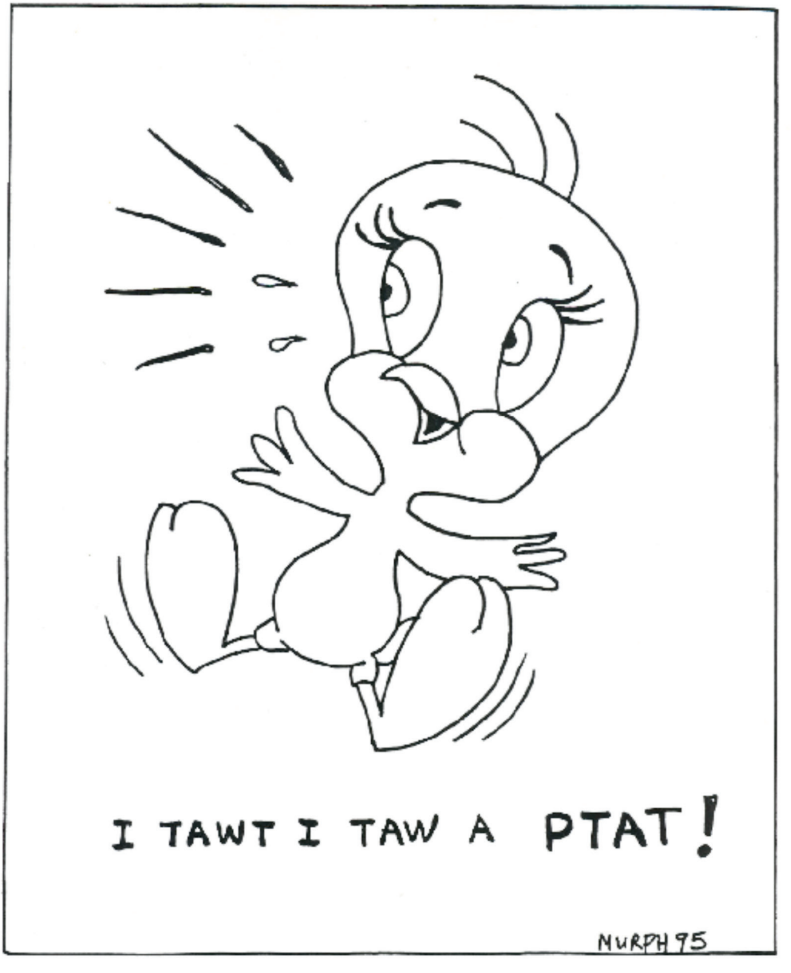
The Bottom Line:

$$\Delta Vbe = (kT/q)\ln((ic1/ic2)(A2/A1))$$

...and if the Currents are Equal:

$$\Delta Vbe = (kT/q)\ln(A2/A1)$$

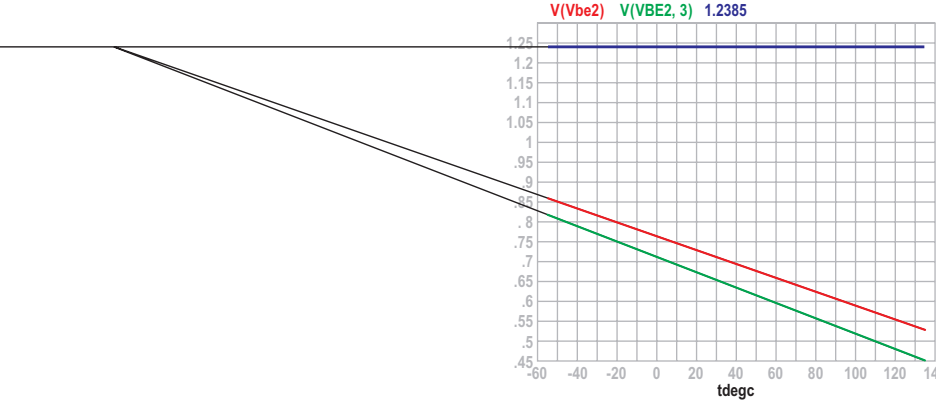
Setting the current ratio equal to one and noting that the area ratio and the constants are unlikely to change, we find that ΔVbe is Proportional To Absolute Temperature …



... or PTAT.

(as it's often referred to)

Current Density vs. Base Voltage Difference

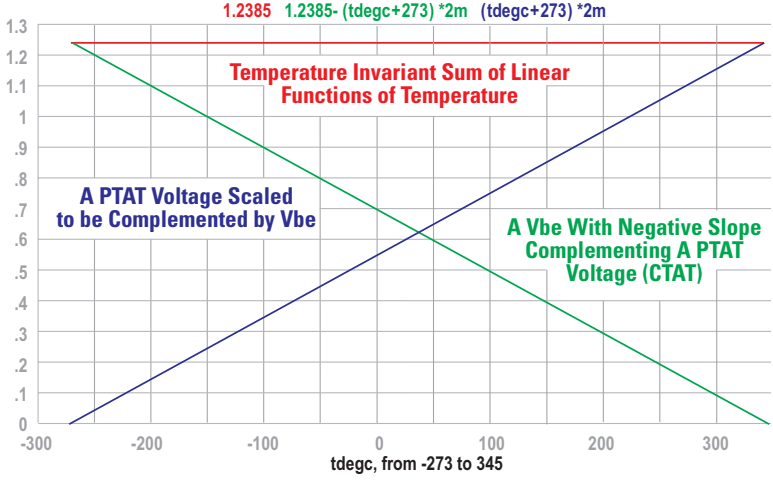


If we extrapolate temperature to zero Kelvins we should find that ΔVbe also goes to zero despite the fact that the transistors may be differently sized and operated at different currents. This figure shows a linear interpolation of Vbe back to zero. Ideally this extrapolation would go to VG0, the extrapolated bandgap of silicon. But the actual value of the two Vbes would meet at a slightly lower voltage than the extrapolation because the slope of Vbe vs. temperature is not quite linear. That is, the linear extrapolation reaches about 1.2385V at zero Kelvin, but tracing out the complete theoretical characteristics of actual transistors would find they meet at a slightly lower voltage. The curvature of Vbe is difficult to see in the figure, unless you look closely to see the departure from the linearity of the extrapolation in the region where the transistors were accurately simulated as indicated by the colored traces.

In practice this curvature over the temperature range of common interest is small enough to often neglect. However, if better temperature behavior is needed several methods may be used to compensate it, one of which will be mentioned later.

For now, let me neglect Vbe curvature and treat it as a linear function of temperature. As we have seen, the PTAT voltage across R1 can be scaled to be larger as it is by R3 (40K) and R4 (40K). So, we can generate a PTAT voltage scaled as we wish to add to the Vbe of Q2. Near our temperature range of interest we should be able to combine Vbe2 with a PTAT voltage scaled to give a wide range of results for the sum. However, as we approach zero Kelvin the PTAT voltage approaches zero, so that whatever Vbe it is combined with will approach VG0, the extrapolated bandgap voltage.

The combined Vbe2 and PTAT voltage, being the sum of two functions linear in temperature, should itself be such a linear function. And since we know one value it will take on, we can make that function constant at all temperatures by adjusting the PTAT voltage to make the sum the same as VG0 at any other temperature, including room temperature. And since these voltages complement one another over temperature Vbe2 is often called Complementary To Absolute Temperature, or CTAT.



Current Density vs. Base Voltage Difference

A simple circuit can be used to generate the sum of both the PTAT and CTAT voltages.

The amplifier used to drive VBE2 until its inputs balance will try to do that even if there is some additional impediment, such as R2, in the way. That is, the amplifier will drive VBE2 positive and the current necessary to drive R3 and R4 will be supplied by a rising voltage on R2.

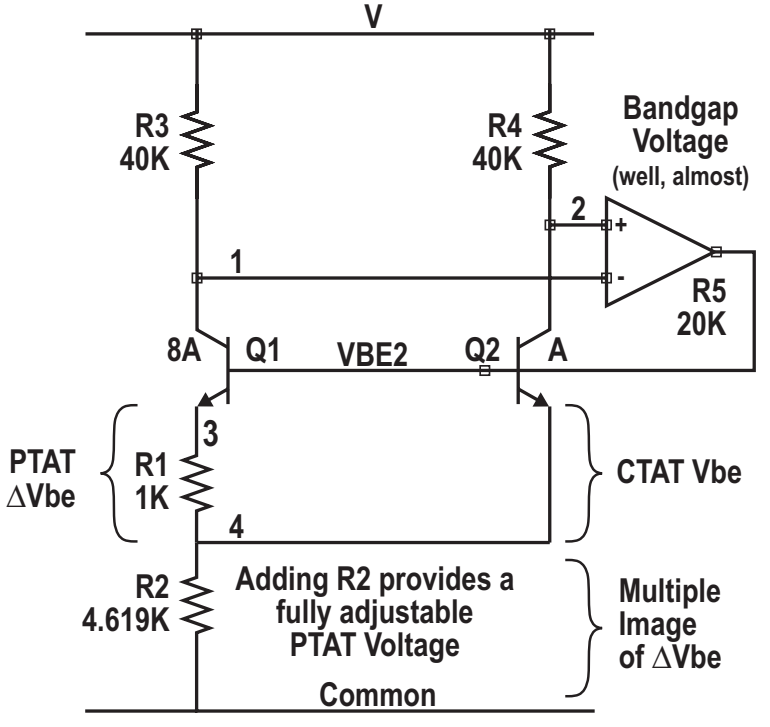
When the two collector currents match, their sum will pass through R2 (actually a bit more when we include the base currents of Q1 and Q2, but they can often be neglected, or made negligible by additional circuitry) making it take on the PTAT voltage characteristic of current in R1.

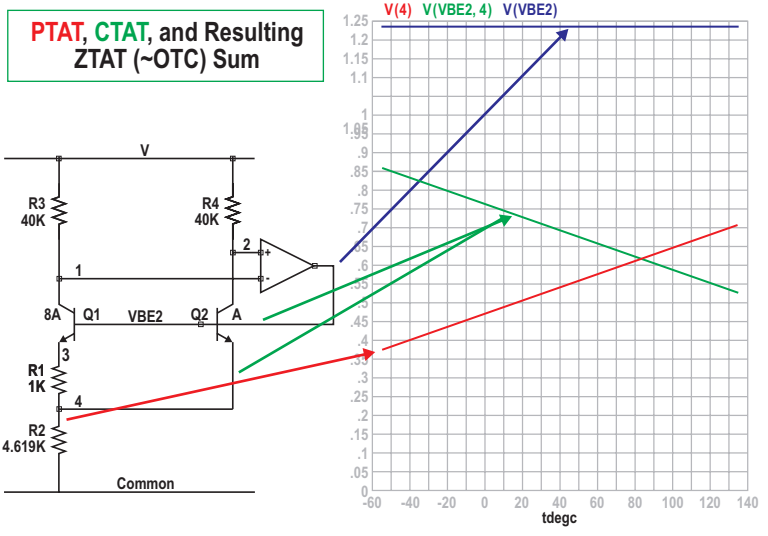
Since we are free to make R2 whatever value we choose, it can be made large enough to be complemented by the CTAT Vbe at the output of the amplifier.

Since the CTAT voltage is not strictly linear, a slight additional PTAT voltage component will tilt up the downward curvature for a best fit to our desired temperature range. This linear addition acts like the extrapolation which, as we said before, passes through a voltage which is a small amount larger than the bandgap.

This small amount will be process dependent and the total is often called the **"Magic Voltage"** for the process. The magic voltage is the output voltage setting at some selected temperature which gives the best temperature performance. Designs to hit the magic voltage can also be trimmed, for example by adjusting R2 at a single temperature.

Disclaimer: This configuration is optimized for presentation, not necessarily for integration

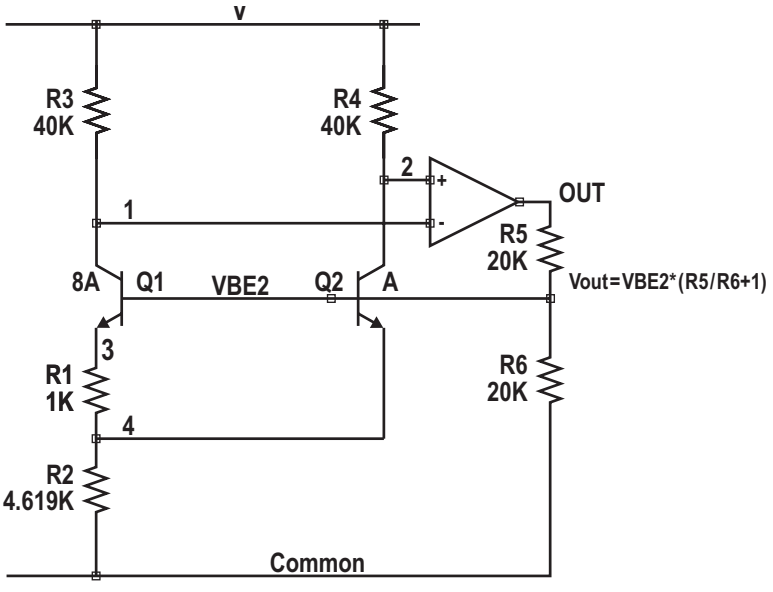




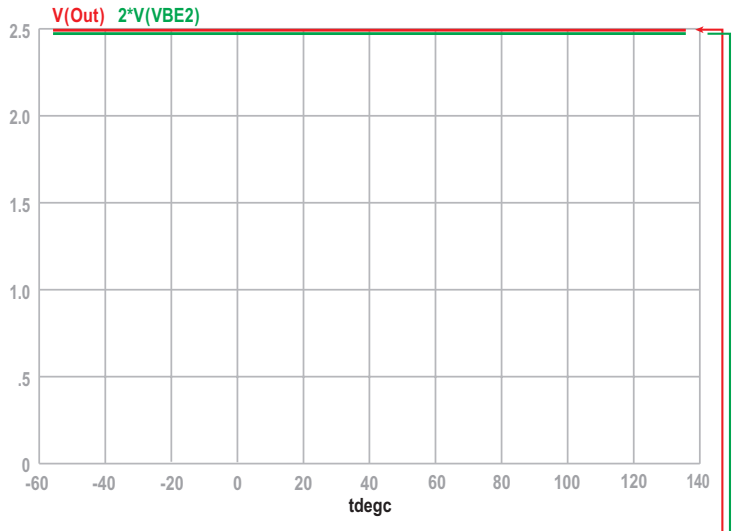
An example of the simulated result for the process used in the other examples shows the low Temperature Coefficient (TC) result at the node VBE2, of adding Vbe2 to the voltage across R2. The apparent TC, at this viewing scale, appears flat and so is often abbreviated ZTAT.

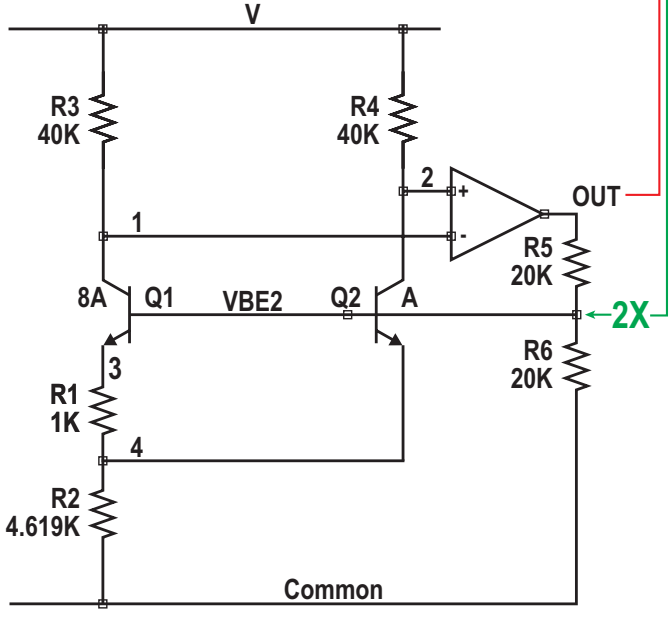
The magic voltage for a given process may be less than some desired reference level such as 1.5V or 2V. And/or it may be desirable to make a design easily reusable in other processes which have different magic numbers.

A Small Additional Burden on the OP-Amp Makes Available Any Desired Output Voltage Greater Than the Bandgap

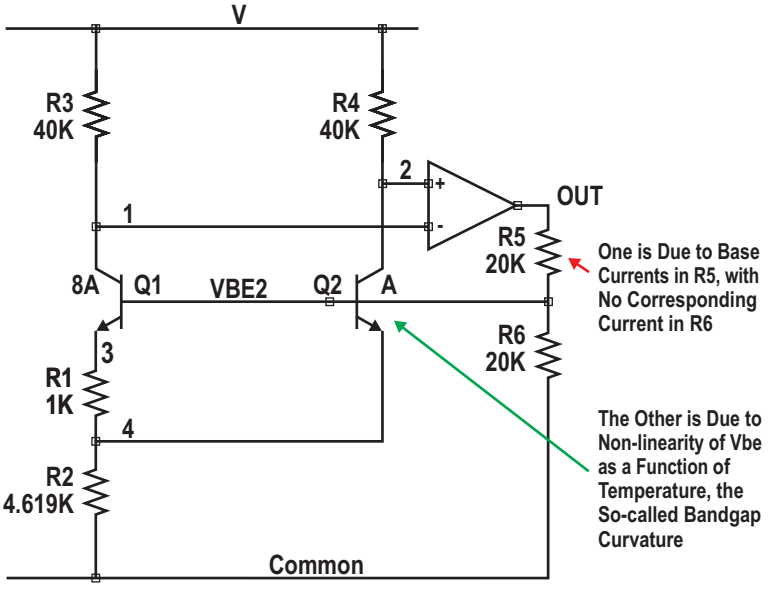


This requirement can often be accommodated with the addition of a simple voltage divider. As we saw before the amplifier will do whatever it can to balance the two collector currents. So, attenuating the output before driving the common base line will require the amplifier output to rise in proportion to the attenuation to restore the magic voltage to VBE2. As shown in the figure, a resistor divider of one half will roughly double the output voltage. The green trace shows twice the magic voltage at VBE2 while the red trace shows the simulated output.


V(Out) 2*V(VBE2)
2.5
2.0
1.5
1.0
.5
0
-60
-40
-20
0
20
40
60
80
100
120
140
tdegc


V
R3
40K
R4
40K
2
+
OUT
1
-
R5
20K
8A
Q1
VBE2
Q2
A
2X
R6
20K
3
R1
1K
4
R2
4.619K
Common

Our Little Circuit Can Reveal Two Fundamental Error Sources

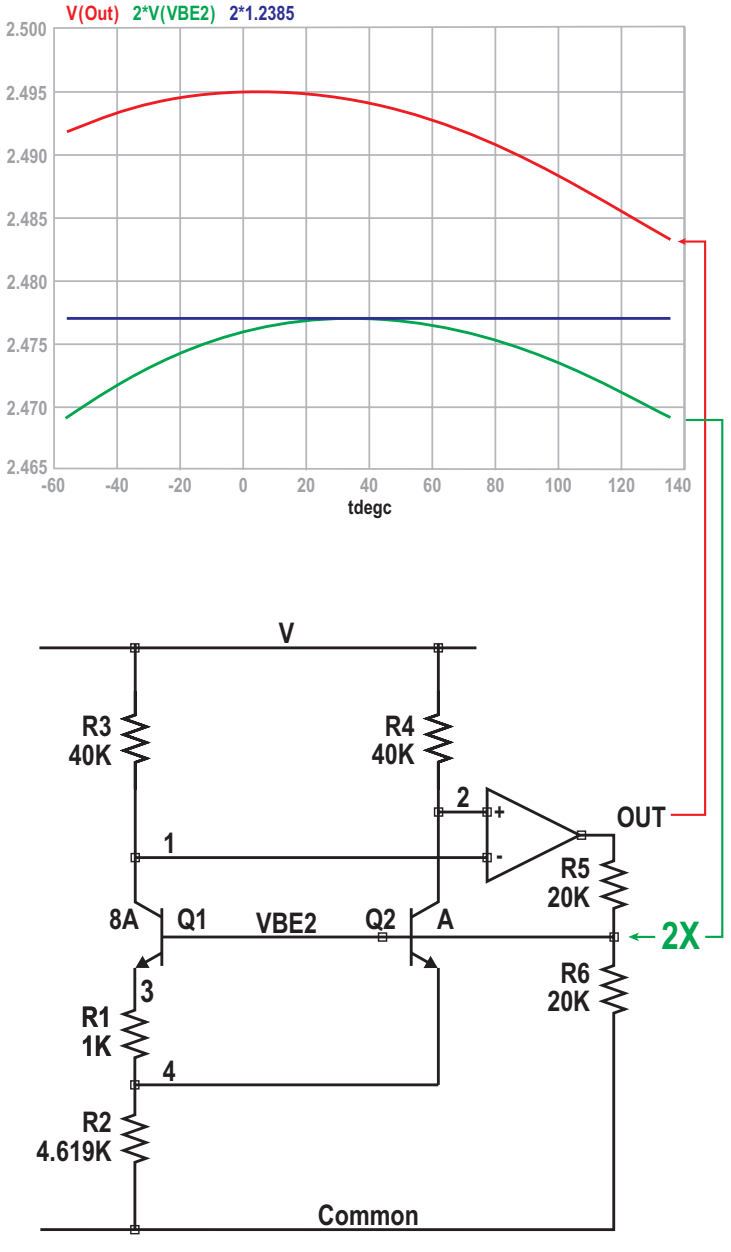


Two errors contributing to non-zero TC can be visualized with this circuit. The effects of base current on the basic reference can often be neglected. However, depending on the current you can afford to spend making the voltage divider from small value resistors, the base current effects on the multiplied output may not be negligible. We can also use this example to investigate curvature compensation.

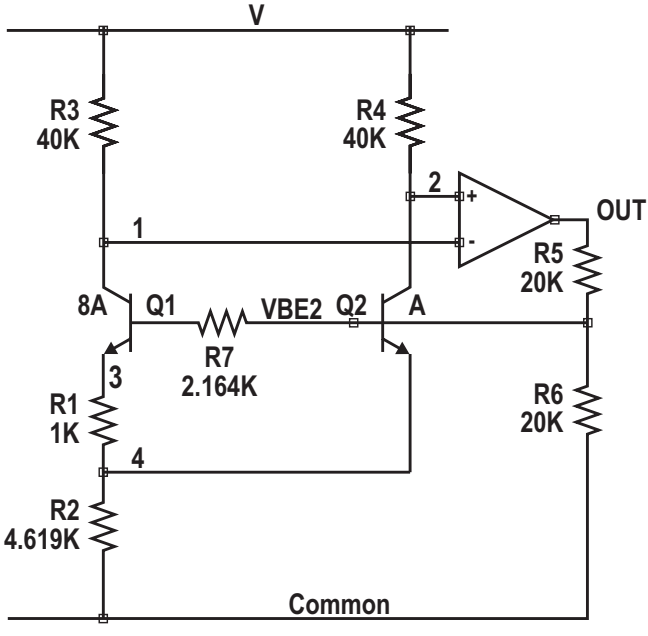
Up until now we have been looking at results on a scale where the total error spanned a pixel or two and looked pretty small. But if we examine the results and the scale of the errors we'll find they look worse and may require reduction to meet our performance specification.

The green trace represents about as well as we can do with the simple cell and the process at hand. It's multiplied by two for comparison with V(Out). A similar arrangement should help you determine both the magic voltage and the intrinsic curvature due to your process. The blue curve is twice the magic voltage, which is ideally tangent to the real or simulated result near the midrange temperature. The red trace is the simulated output with a 2X feedback attenuation.

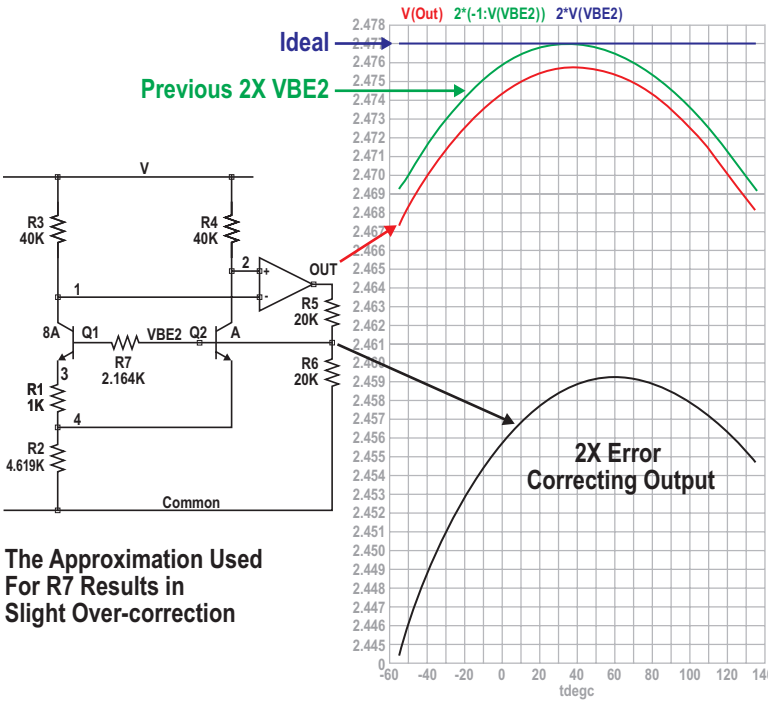
The reason it is higher than the others is that the base currents of both transistors flow in R5, slightly raising its voltage, without a corresponding component of current in R6. Neglecting a few things, like the possibility that the base currents differ because of collector current density, we can make an approximate correction for this error.


V(Out) 2*V(VBE2) 2*1.2385
2.500
2.495
2.490
2.485
2.480
2.475
2.470
2.465
-60 -40 -20 0 20 40 60 80 100 120 140
tdegc
V
R3
40K
R4
40K
2
+
-
OUT
1
R5
20K
8A
Q1
VBE2
Q2
A
2X
R6
20K
3
R1
1K
4
R2
4.619K
Common

R7 introduced between Q1 and Q2 carries the base current of only Q1 and introduces a slight offset into the loop, making ΔVbe, which changes the current in R1 and, hence in R2. The formula given in the figure gives a good first approximation which can be refined with a bit of experimentation.



Adding R7=(R1/R2)*R5*R6/(R5+R6) creates a base current proportional voltage within the ΔVbe Loop, reducing the reference voltage, slightly, to compensate the error due to base currents in R5



The Approximation Used For R7 Results in Slight Over-correction

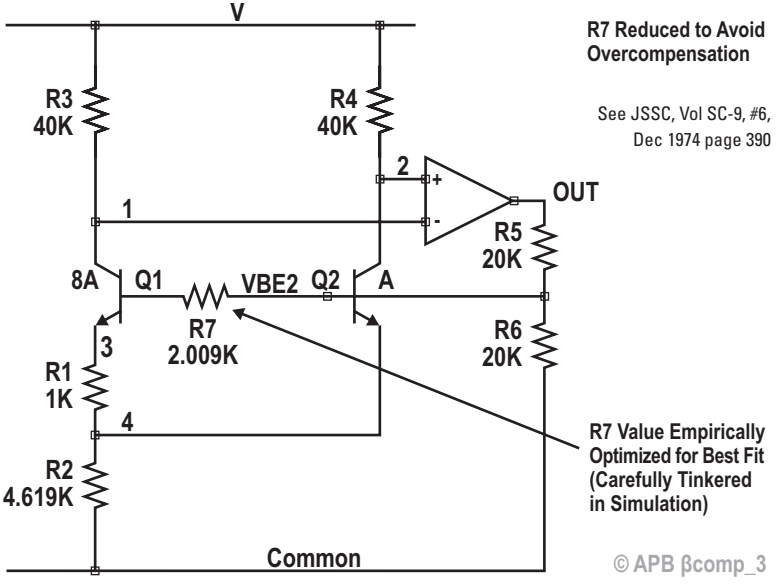
Using the formula with the schematic values gives a value for R7. The simulation in black shows the effect of R7 and base current on the nominal voltage at VBE2.

The red trace shows the simulated output of the circuit to compare with twice the value of VBE2 from the previous simulation.

The black trace shows twice the pre-compensated value to compare with the simulated output which results from base current in R5.

The simulated output is a bit less than twice the nominal magic voltage, since the approximation of the formula slightly overcompensated the base current.

R7 can be reduced slightly to take into account a well-modeled difference in the base currents of Q1 and Q2 as well as the approximations in the formula for R7.

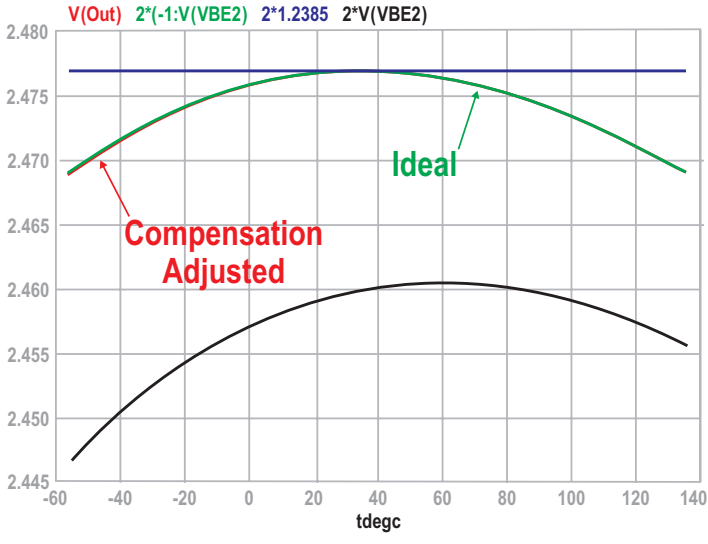


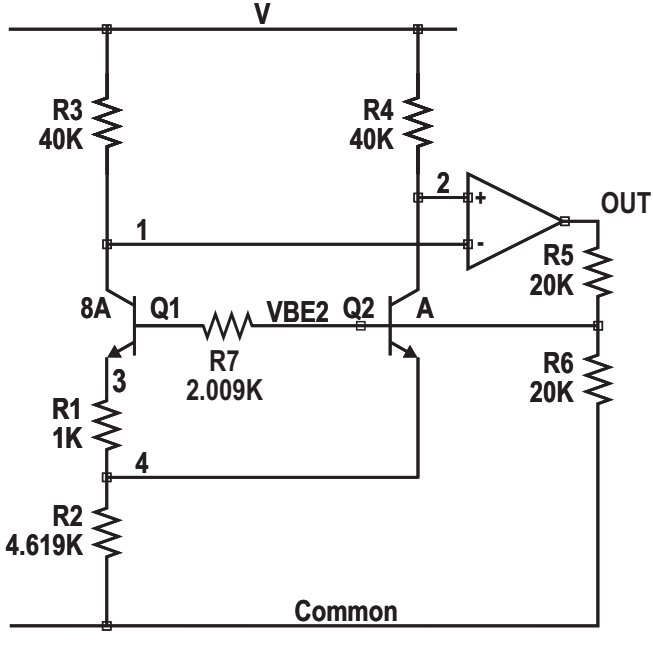
The result of this adjustment can be seen here to almost exactly match the simulated output to an ideal doubling of the output of the single bandgap, indicating that the output error due to base currents has been greatly reduced.

The pre-distorted voltage at VBE2 is also doubled and shown in black to compare with the ideal doubling.

As caveat, I want to point out that while this error reduction technique works well in practice you may prefer to find an alternative if noise at the output is an issue. The addition of R7 can easily double the intrinsic minimum noise of this circuit and must be considered along with other substantial noise sources.

All of the outputs we have seen show a non-linear voltage vs. temperature component. Unlike PTAT errors, which can be corrected with PTAT voltages that are easily generated, the PTAT output adjustment can only correct for the linear baseline of the nonlinearity resulting from Vbe.


V(Out) 2*(-1:V(VBE2) 2*1.2385 2*V(VBE2)
2.480
2.475
2.470
2.465
2.460
2.455
2.450
2.445
-60 -40 -20 0 20 40 60 80 100 120 140
tdegc
Ideal
Compensation
Adjusted


V
R3
40K
R4
40K
2
+
-
OUT
1
R5
20K
8A
Q1
VBE2
Q2
A
R7
2.009K
3
R1
1K
R6
20K
4
R2
4.619K
Common

So what about the Curvature?

Where does THAT come from??

The origin of the curvature is with Vbe, as is almost everything in this band-gap circuit. An expression for Vbe derived from first principles allows us to compare Vbe of a transistor with a second transistor which matches it or with itself when operated under different conditions. For a given process and device design, then, the expression for Vbe lets us say "If you've seen one, you've seen 'em all."

Vbe of a Transistor as a function of Current and Temperature may be inferred from a measurement of Vbeo, its nominal value at temperature T=To, and current i=io, using the relation:

$$Vbe = VG0 + (T/To)(Vbeo - VG0) + (kT/q)\ln(i/io) + m(kT/q)\ln(To/T)$$

Where **VG0** is the Bandgap Voltage extrapolated to 0K and **m** is a property of the Transistor determined by its design and processing. **m** is approximated by **XTI** in the GP SPICE Model

Note that **Vbe** Consists of:

- a Constant
- a Term Linear in Temperature
- a Term that May be Zero or nonlinear
- a Term which is a non-zero, non-linear, function of temperature

This last is the origin of the "Bandgap Curvature."

Since **Vbeo** will be less than **VG0**, the dominant linear term will fall with increasing temperature ... And, the third term, **(kT/q)ln(i/io)**, will be zero, if the current is held invariant at io.

However — if **i=io**(T/To), as is the PTAT current in our bandgap cell, then this term may be re-written as:

$$(kT/q)\ln(io(T/To)/io)) = -(kT/q)\ln(To/T)$$

which may be combined with the fourth term,
which is the curvature term, to yield:

$$(m-1)(kT/q)\ln(To/T)$$

The Vbe of Q2 Operating with PTAT Current is combined with the voltage across R2 scaled up from ΔVbe to produce the nominal output voltage:

$$Vout = VG0 + (T/To)(Vbeo-VG0) + (m-1)(kT/q)\ln(To/T) + R2((kT/q)\ln 8)/R1$$

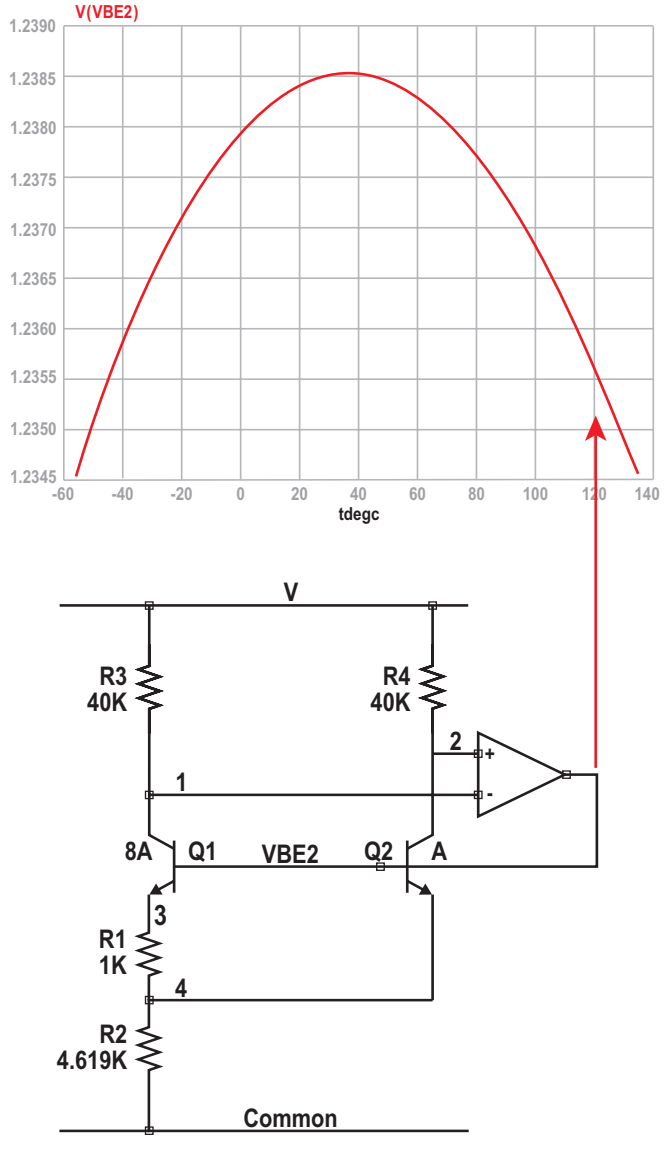
The second and fourth terms are both temperature proportional and are of opposite sign, so their sum may be made zero by setting the R2/R1 ratio. This leaves **Vout** as:

$$Vout = VG0 + (m-1)(kT/q)\ln(To/T)$$

VG0 is the ideal bandgap voltage of silicon extrapolated to zero, and would be a fine reference value if the total was not corrupted by the second, curvature, term.

This term is generally curving downward in the temperature range of interest to us so we add a PTAT component (by adjusting R2/R1 to leave a residual + PTAT) to remove the net negative slope of the curvature. As a result we see the peak of the curvature in our temperature range, slightly above the ideal VG0.

A Diferent Perspective on the Basic Circuit Using a Different Scale



Going back to the basic bandgap circuit and looking at the output on a different scale shows the curvature as if superimposed on our desired temperature invariant circuit. Although the curvature term is a logarithmic function of temperature in a series expansion it has a large quadratic term, for which we can compensate and reduce the remainder to a third and higher order error.

The way in which the current variable term, (kT/q)ln(i/io), reduces the final curvature term may be exploited by generating a larger multiple to combine with the fundamental curvature of the same form to reduce the sum to zero.

A more simple and cost effective compensating curvature can be made by adding a resistor with a more positive TC to the PTAT voltage multiplier.

Note that the residual curvature appears to be a parabola, indicating that it has a large quadratic component.

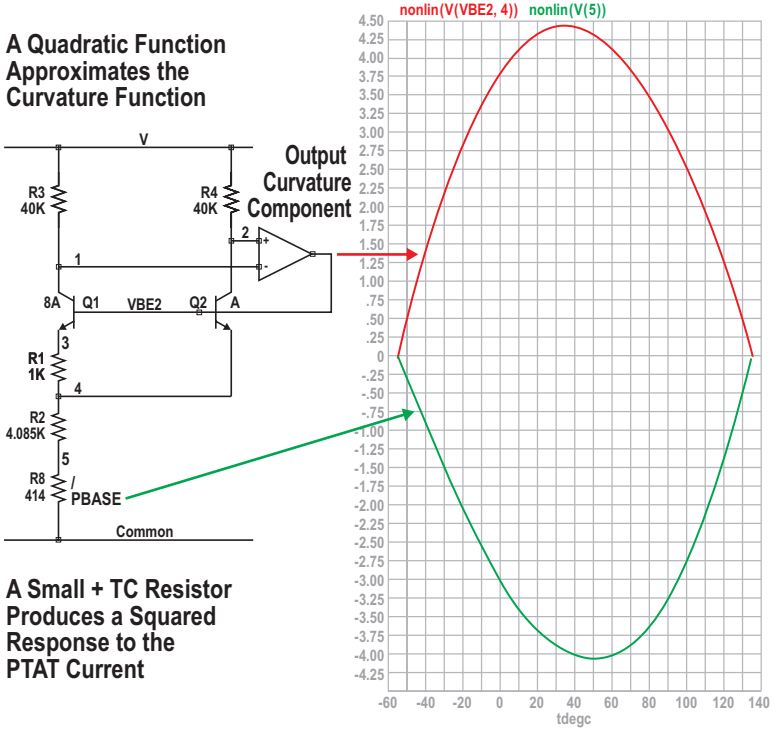
A temperature proportional resistor driven by the PTAT current results in a T2 voltage which can be sized to cancel the quadratic component of the residual curvature.

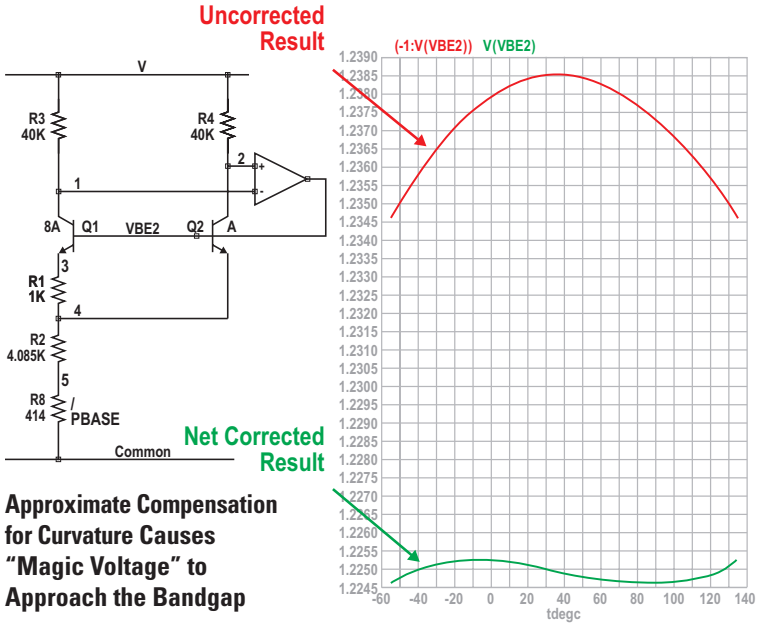
The small remainder will be seen to be dominated by a third order term.

If we put a PTAT current in a temperature proportional resistor, the result is a quadratic function of temperature. We can do that by putting a diffused resistor with a, relative to R1 and R2, large positive TC in series with R2. Of course R2 must be reduced, not only to "make room" for the R8 voltage, but to also reduce the total voltage across R2 and R8. As we will see, curvature correction gives us better performance, and also brings the magic voltage closer to the actual extrapolated bandgap.

This figure hides the zero TC component of the previous plotted output voltage to show the curvature due to Vbe of Q2. Over the same temperatures we can see the quadratic compensating voltage appearing across R8, an added +TC resistor.

The summation of the quadratic correction with Vbe curvature substantially reduces the remaining temperature error.





Approximate Compensation for Curvature Causes "Magic Voltage" to Approach the Bandgap

Note that the magic voltage that results from curvature correction is substantially reduced, making it closer to VG0 where it would be without the effects of Vbe curvature.

So far, the example circuits and simulations have been based on the SPICE models of an analog bipolar process. Although the Bandgap Reference principles were developed using similar transistors, the need for voltage references persists on circuits made on processes that lack isolated vertical transistors.

... and Now, for the Device Deprived ...

From time-to-time some of us may find ourselves Deprived of real (vertical, isolated, bipolar) transistors and forced to make do with crumbs from the CMOS table.

Fortunately, even parasitic vertical substrate transistors will exhibit the

$$\Delta Vbe = (kT/q)ln((ic1/ic2)(A2/A1))$$ relationship,

and the Emitter Current ratio MAY be a satisfactory substitute for (ic1/ic2)

Moreover, these ill suited thick-base substitutes may be more uniform in manufacture than their thin-base counterparts that serve us so well. This uniformity is welcome, though it hardly compensates for the lack of individually usable collector currents.

At the heart of the bandgap reference is a junction, or two or three. On CMOS processes, diffusions may be arranged to form transistors that have a common collector in the substrate. The opportunity to directly control the collector current ratio by comparing them is not available. Additionally, since the collector currents are not driving a second stage, ΔVbe must be measured directly rather than inferred from a higher gain common base input stage.

A CMOS Bandgap Reference Circuit with Sub-1-V Operation

Hironori Banba, Hitoshi Shiga, Akira Umezawa, Takeshi Miyaba, Toru Tanzawa, Shigeru Atsumi, and Koji Sakui, *Member, IEEE*

***Abstract*— This paper proposes a CMOS bandgap reference (BGR) circuit, which can successfully operate with sub-1-V supply. In the conventional BGR circuit, the output voltage V_{ref} is the sum of the built-in voltage of the diode V_f and the thermal voltage V_T of kT/q multiplied by a constant. Therefore, V_{ref} is about 1.25 V, which limits a low supply-voltage operation below 1 V. Conversely, in the proposed BGR circuit, V_{ref} has been converted from the sum of two currents; one is proportional to V_f and the other is proportional to V_T. An experimental BGR circuit, which is simply composed of a CMOS op-amp, diodes, and resistors, has been fabricated in a conventional 0.4-μm flash memory process. Measured V_{ref} is 518 ± 15 mV (3σ) for 23 samples on the same wafer at 27–125°C.**

***Index Terms*—Bandgap reference, CMOS, low voltage.**

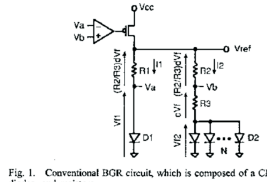


Fig. 1. Conventional BGR circuit, which is composed of a CMOS op-amp, diodes, and resistors.

I. Introduction

REFERENCE voltage generators are used in DRAM's, flash memories, and analog devices. The generators are required to be stabilized over process, voltage, and temperature variations, and also to be implemented without modification of fabrication process. The bandgap reference (BGR) is one of the ... voltage generators that succeeds ...

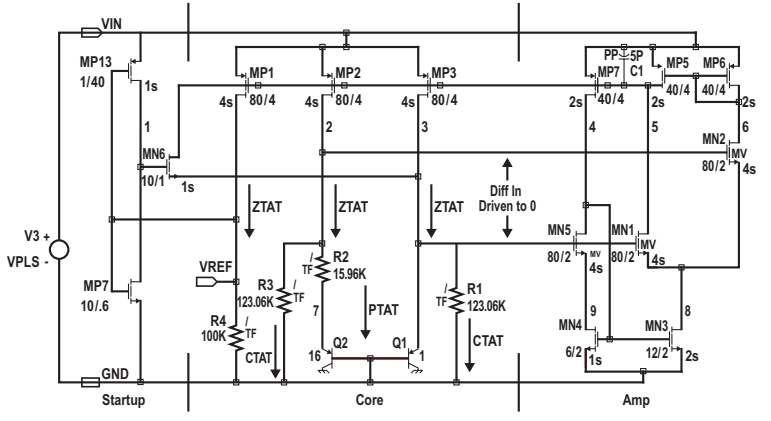
is expressed as

$$I = I_s \cdot \left(e^{q \cdot V_f / k \cdot T} - 1\right) \cong I_s \cdot e^{q \cdot V_f / k \cdot T} \quad \left| V_f \gg \frac{k \cdot T}{q} \right. \tag{1}$$

$$\ldots = V_T \cdot \ln \ldots$$

This paper shows in some detail a way to build a temperature stable reference using parasitic Vertical PNP transistors. And to generate an output voltage which may be freely chosen within a useful range. A similar and instructive circuit can be used for illustration.

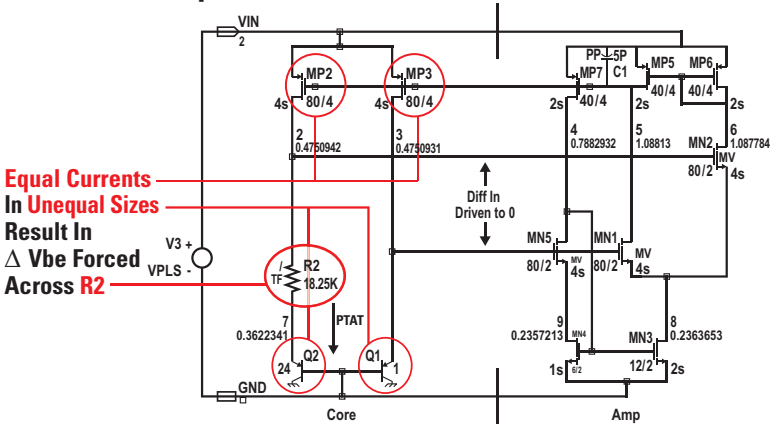
Simple sub-Bandgap uses CMOS Parasitic PNPs



This figure shows a relatively simple scheme to derive a 1V reference using VPNP transistors Q1 and Q2. They are part of the core, the center section of the diagram, which is flanked by an Amplifier and a Startup circuit.

To simplify the explanation, we can strip away the Startup and part of the core.

But First, A Simplified Sub-Block



The Amplifier, at the right, drives the gates of two matched PMOS transistors MP2 and MP3, causing them to deliver equal currents to Q2 and Q1 which are in the ratio 24:1 of matched emitters. As long as the currents in the two transistors are equal and non-zero the difference in their Vbes should be given by:

$$\Delta Vbe = (kT/q)\ln 24$$

depending only on the ratio of the currents, not their magnitude.

That means that when the matched currents are small the emitter voltage of Q1 is larger than the emitter voltage of Q2 by almost ΔVbe. This will make the gate voltage of MN1 more positive than the gate of MN2 and the drain of MN1 will pull down the gate of MP3 as well as of MP2 connected to node 5.

Increasing MP3 current will, of course, increase Vbe of Q1 as a positive feedback.

However, as currents increase, so does the voltage across R2 until it eventually is equal to ΔVbe. At that point, the difference between the gate voltages of MN1 and MN2, the inputs to the amplifier, will be zero. Considering the gain of the path from MP3 gate (node 5) to node 1 at the emitter of Q1 it will be limited by the transconductance of MP3 and the impedance of Vbe1.

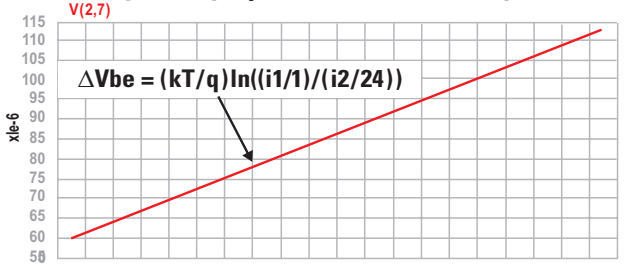
The gain of the path from MP2 gate to node 2 at the top of R2 will be limited by MP2 Gm and the combined impedance of R2 with Q2. Since the impedance of Q1 and Q2 are the same at equal currents, the negative feedback path will dominate the positive feedback path and the circuit will rest at equilibrium with currents set by ΔVbe/R2.

Simulation over temperature shows the PTAT voltage appearing across R2.

The second trace is the input voltage of the amplifier which is within 1uV over temperature. The amplifier topology yields these results, but of course is corrupted by device matching errors.

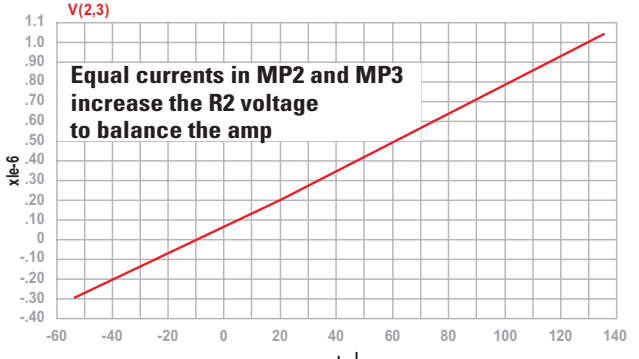
We can be certain that when the MP3 and MP2 currents are equal, the amplifier inputs will be balanced and bring their drains to the voltage required to set up the PTAT currents.

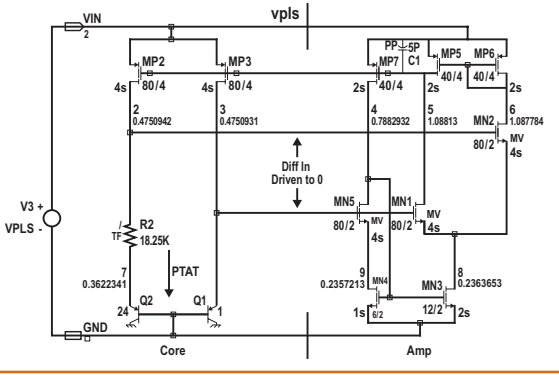
Developing a PTAT voltage across R2 by forcing equal Node 2 and 3 voltages



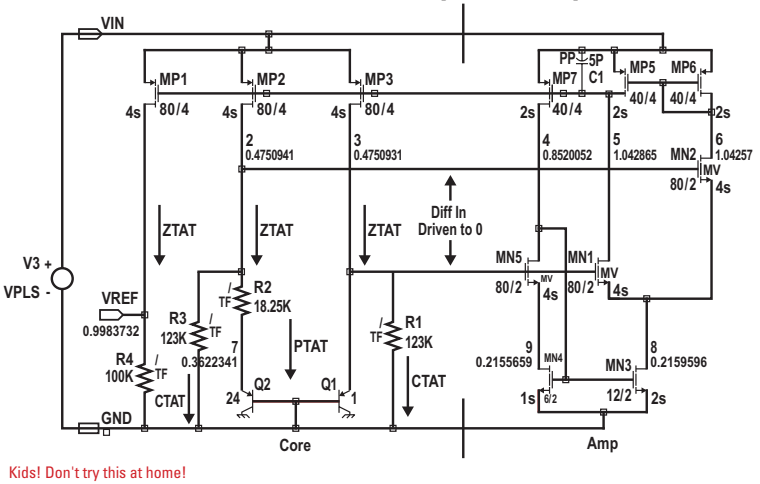
Q1 and Q2 operate at equal currents set by MP2 and MP3, but they are sized 1:24, resulting in a 24X current density ratio and a difference in Vbe given by: **ΔVbe = (kT/q)ln((i1/1)/(i2/24))**

The amp forces ΔVbe to appear across R2





Loading Nodes 2 and 3 requires the Amp to make more MP2 and MP3 Current to preserve equilibrium



Kids! Don't try this at home!
(It's a non-starter)

If we load, equally, nodes 2 and 3 the amplifier must drive MP2 and MP3 to supply this additional current in order to maintain the input voltages which force the PTAT current.

Equal valued R1 and R3 added to the schematic will require more current from MP2 and MP3, and the amplifier will drive them to provide it.

Notice that the currents in Q1 and Q2 should remain PTAT while the currents in R1 and R3 are proportional to Vbe and must, therefore, be CTAT. MP2 and MP3 are supplying both. Remembering that the sum of a CTAT and PTAT quantity can be made temperature invariant by using the correct proportions, we can adjust either or both of R2 with R1and R3 to make a temperature invariant current in MP2 and MP3. Such a quantity is sometimes referred to as ZTAT, in this context.

The ZTAT current should also flow in MP1, a transistor matched to MP2. This current can drive the load resistor, R4, to a voltage resulting from the ZTAT current flowing in it.

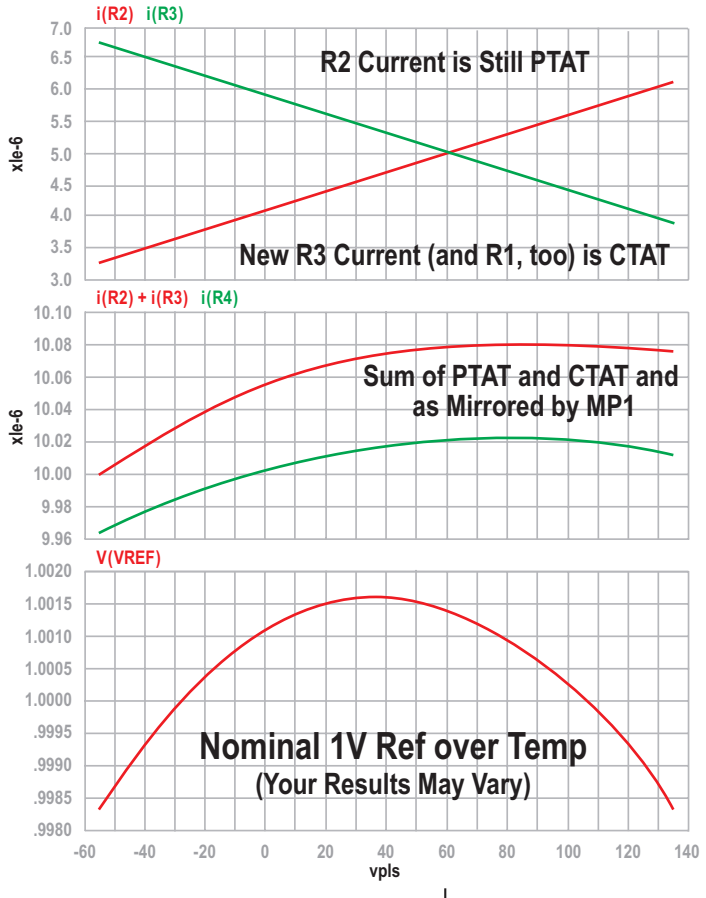
Loading CTAT Voltages With R1 and R2 Requires Complementary Currents From MP2 and MP3

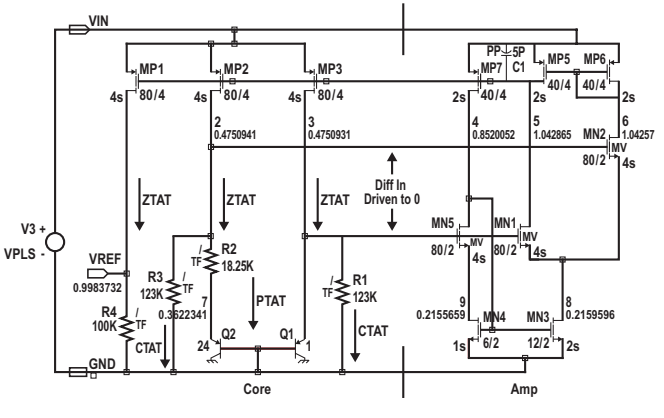
This figure illustrates the PTAT and CTAT currents and their almost constant sum at scale that reveals an error and a consequence. The error is due to the curvature of Vbe vs. temperature, and will remain in the final output. The consequence is a slight net positive TC to the current, due to a small TC in the thin film resistors used for simulation. This TC should not appear in the output, since R4 has the same TC.

The output voltage results from approximately 10μA flowing in R4, so that it can be set over a wide range by changing the nominal value of R4.

The small difference seen between i(R4) and i(R2) + i(R3) is mostly a result of the difference in the drain voltage of MP1 from that of MP2 and MP3. In many applications this may be neglected, or if Vin has a wide range of values, the three ZTAT transistors can be cascoded to reduce supply voltage sensitivity.

This circuit shares a common "gotcha" with most self biasing circuits. In addition to the desired behavior described here, the circuit can take on another stable state when all currents and the reference voltage are at zero. Once in this state the circuit will remain there.


i(R2) i(R3)
R2 Current is Still PTAT
New R3 Current (and R1, too) is CTAT
xle-6
i(R2) + i(R3) i(R4)
Sum of PTAT and CTAT and as Mirrored by MP1
V(VREF)
Nominal 1V Ref over Temp
(Your Results May Vary)
vpls


VIN
MP1
MP2
MP3
MP5
MP6
MP7
C1
5P
ZTAT
Diff In
Driven to 0
V3 +
VPLS -
VREF
0.9983732
R3
123K
R4
100K
R2
18.25K
R1
123K
PTAT
CTAT
Q1
Q2
GND
MN1
MN2
MN3
MN4
MN5
MV
Core
Amp

So, most self-biased circuits require a startup which is sometimes easily arranged, other times not. My personal preference is to avoid dynamic starts, since once started and switched off, for example by noise, your circuit may then fail to restart.

This circuit includes a starter circuit that operates when power is applied to VIN.

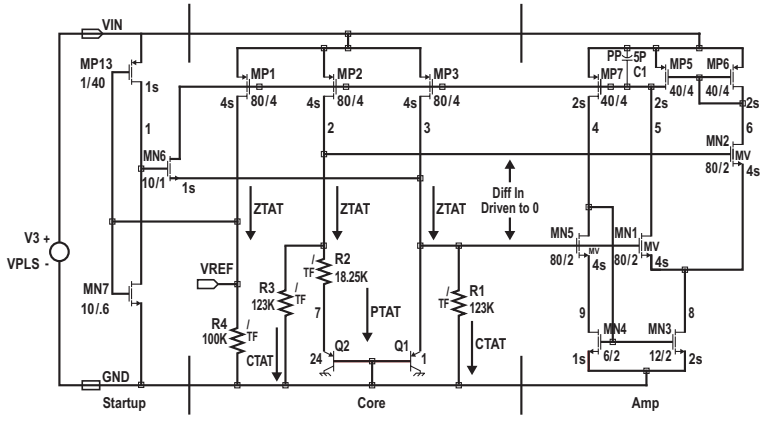
If the circuit does dynamically start (depending on the dV/dt of VIN among other things), fine, but if not, VREF will be low ~ 0. As a result, MP13 will come on as power ramps up. It will turn on M6 which drives the common PMOS gate line near the top turning on MP2, 3, and-so-forth. Once there are some non-zero currents the circuit will come on regeneratively.

The current in MN6 is likely to be very much larger than the nominal current provided by MN1 to node 5, and it is important to prevent it from corrupting the reference. As the circuit starts and VREF rises, it will drive MN7 on. Note that while MN7 has a relatively large gate aspect ratio, MP11 has a small one and is easily overcome by MN7. When node one is pulled low, MN6 will be turned off and the small current from MP13 will continue to flow in MN7.

This arrangement works well and can be used in other applications. But, a word of warning: The startup circuit works by triggering the positive feedback in the core and amplifier loop. However, the startup circuit includes another negative feedback path in addition to the one that stabilizes the bandgap output. Normally this path is unstable and gives control to the desired loop. However, it may be stabilized by the addition of a low frequency dominant pole, such as one formed by a large capacitor from VREF to GND that might be added to reduce noise. That can cause the startup to stabilize VREF at a voltage lower than the main loop.

But, assuming that problem is avoided, the startup can be observed in a sweep of VIN.

Origin of the expression "Awakened with a Start"

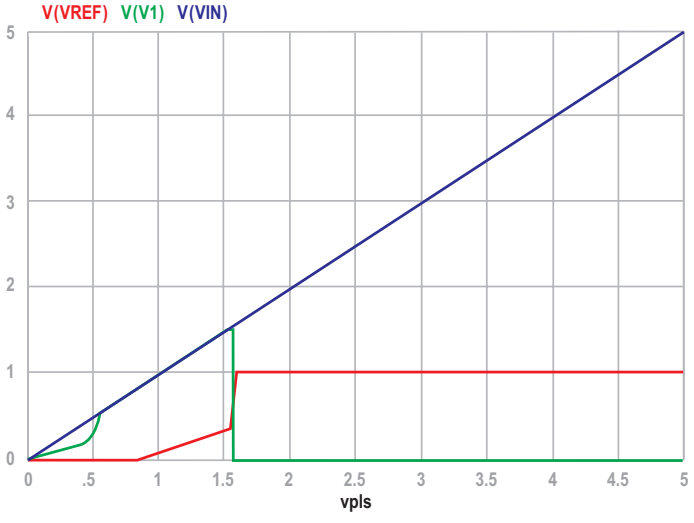


Initially R4 holds VREF low for any small VIN voltage. When VIN is below the threshold of any MOS device, they must all be off or operating sub-threshold at low currents. The green trace shows the simulator's guess at the node 1 voltage for this condition. As VIN reaches the threshold voltage of MP13 it comes on and pulls node 1 to VIN. Initially this voltage is insufficient to turn on MN6, but as VIN continues to rise MN6 connects node 5 (the common PMOS gate line) to node 1 at Q1.

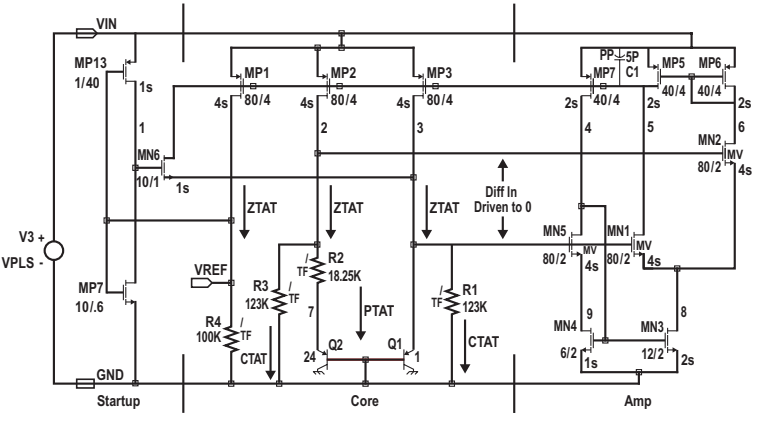
VIN continues to rise so that MP2 and MP3 start to turn on the core, and when VIN reaches about 1.6V the positive feedback of the main loop comes on and drives VREF to the stable operating point. As VREF rises it turns on MN7 which pulls node 1 low and cuts off MN6.

This circuit using VPNPs does not explicitly express the bandgap voltage, although other arrangements can be made to do so. A simple thought experiment may give a different perspective on how the bandgap still remains the reference. At zero Kelvin we should not expect the circuit to work correctly, but we can extrapolate room temperature behavior to that temperature.

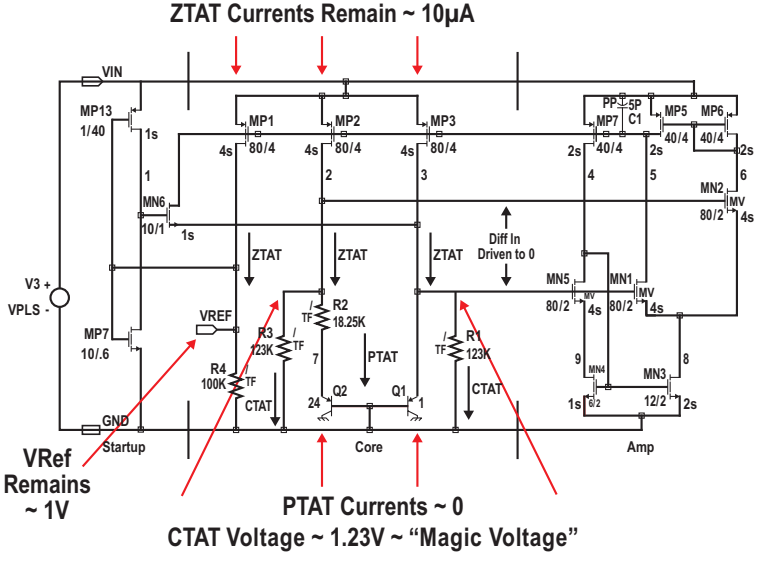
Start-up and Regulation vs. Supply Voltage



MN6 energizes the PMOS Gate Line until VIN provides headroom to start. Then VREF turns on MN7 to sink small drain current of Long Channel MP13



In a Thought Experiment Extrapolating to 0K



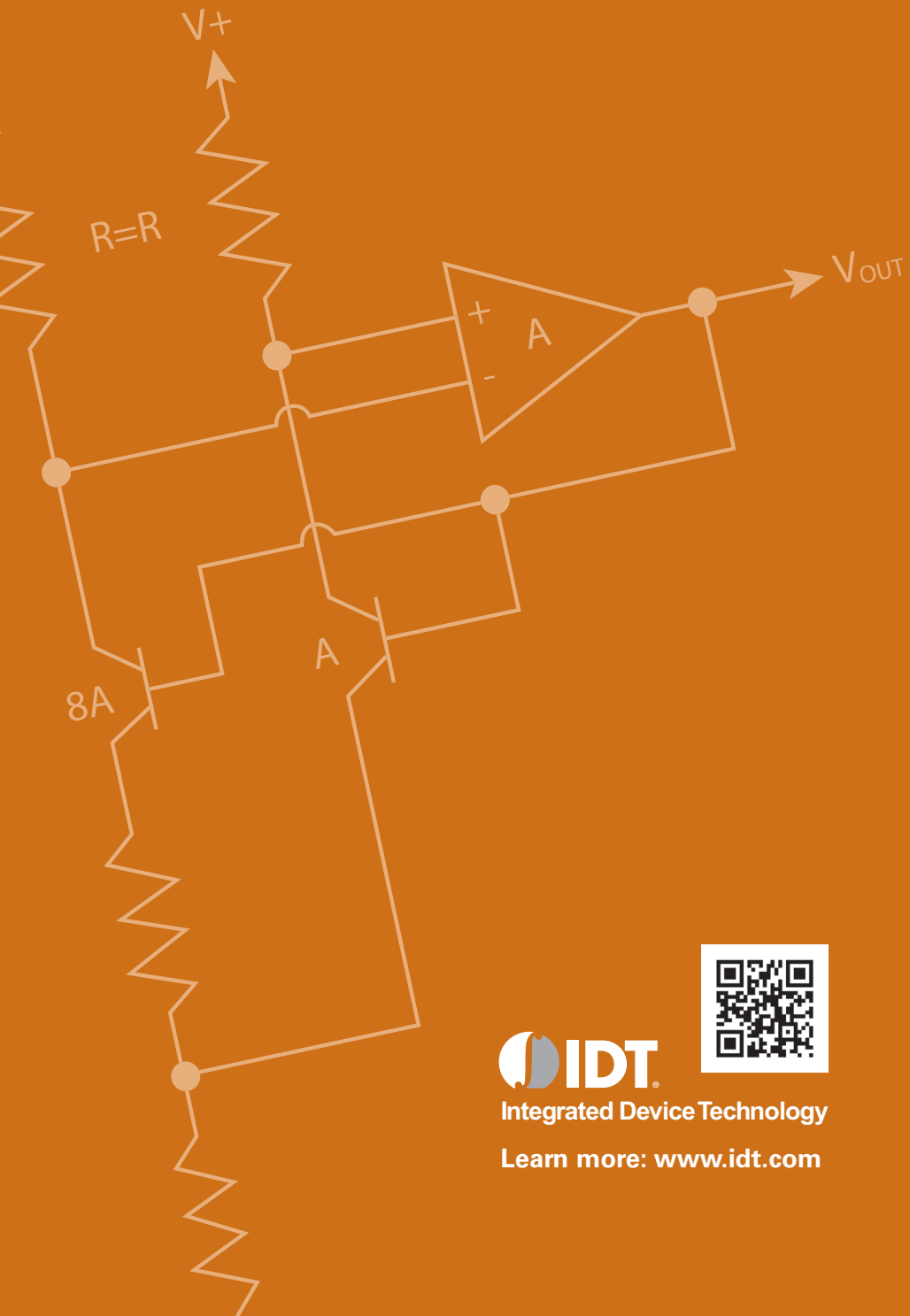
Suppose we take this circuit and imagine the linear approximation of the PTAT and CTAT voltages across R2 and R1, R3, respectively, will continue to zero Kelvin. As we approach zero the PTAT component will approach zero volts, while Vbe should approach the linear extrapolation to VG0. If we also imagine that the ZTAT current will remain invariant at 10µA, all of this current must now be in each of R1 and R3. Using the values which gave the "correct" ZTAT current at room temperature we see that the 10µA in 123K indicates the Magic Voltage should be about 1.23V, for these transistors.

I hope this presentation has been useful and answered a few questions – and I'd like to think it may have evoked even more questions.

We have some of the answers at IDT – perhaps you could help us find more.

V+
V+
R=R
+
A
-
A
8A
VIN
MP13
1/40
1s
1
MN6
10/1
V3 +
VPLS -
MP7
10/.6
GND
Startup

VOUT
MP1
MP2
MP3
4s
80/4
4s
80/4
4s
80/4
PP
5P
MP7
C1
2s
40/4
2
3
4
1s
ZTAT
ZTAT
ZTAT
Diff In
Driven to 0
VREF
R2
18.25K
TF
R3
123K
TF
R4
100K
TF
7
PTAT
CTAT
24
Q2
Q1
1
R1
123K
TF
CTAT
Amp Trick
MN5
MN1
80/2
MV
4s
80/2
MV
4s
9
MN4
MN3
1s
6/2
12/2
Core
Amp



IDT

Integrated Device Technology

Learn more: www.idt.com