Reference generators part 2

August 29, 2019

In the first part dealing with reference generators I showed the Widlar bandgap and the Brokaw bandgap. Both of them use NPN transistors as reference elements. In most CMOS processes these NPN transistors aren't available. So CMOS technologies require a different bandgap design.

1 Bandgap with PNP transistors and parallel resistors

Most CMOS process variants offer a substrate PNP transistor that can be used to build a bandgap. The emitter of this transistor is the drai and source doping of the PMOS transistor. The nwell acts as the base. The P-substrate acts as the collector of the PNP transistor.

The following bandgap is nice and cheap BUT start up is unreliable. Therefore it can be found in hundreds of books but it is not used in too many real chip designs. Nevertheless project managers love it to tell designers they are wasting space using something else.....

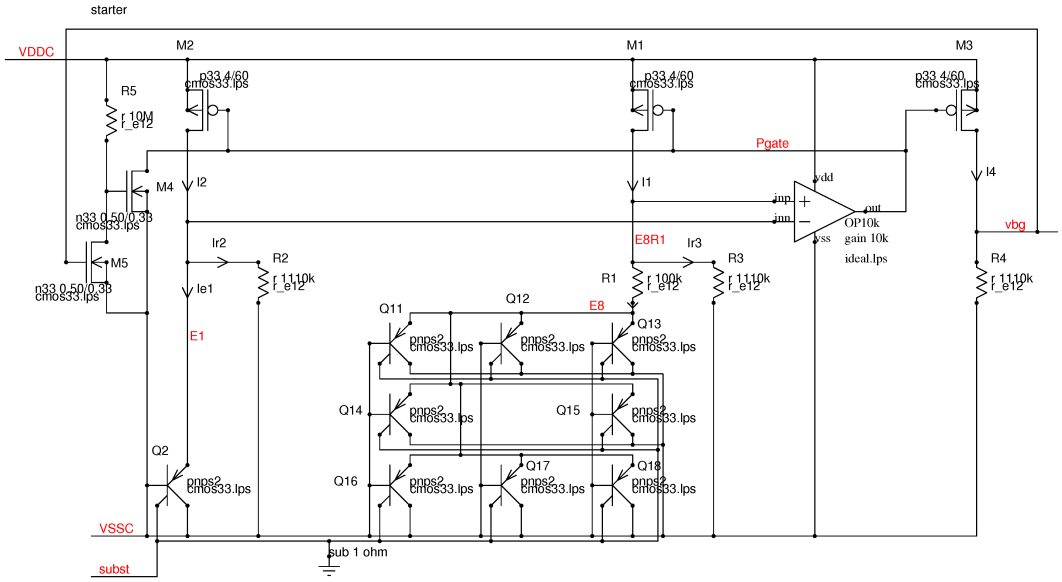

Fig.1.1 The probably cheapest CMOS bandgap

In most CMOS processes the only bipolar component available is the substrate PNP transistor. Sad but true the collector usually is not available for circuit design. It is coincident with the substrate node. So the only thing to be done is to connect the base to ground (metallic circuit ground, reference ground, analog ground or whatever else the top level designer called the ground all precision stages refer to). The emitters are freely available. So the resistors can be connected to the emitters. The temperature voltage is the difference between the single emitter transistor and the multi emitter transistor.

Practical consideration for design anarchists: Your technology does not have this bipolar transistor? Don't worry. Look at the cross sections of the PMOS transistors. Just misuse the PMOS transistors et voila, you can build a bandgap. If you can simulate it depends on the detail level of the model of the PMOS transistors. And if you can't simulate it do it like in the old days. Forget simulation. Use a spread sheet. (The project manager will hate you for this because it is a significant increase of the project risk - but sometimes analog designers have to use methods not foreseen by the design flow.)

So how does that beast work? Assuming M1 and M2 are matching (quite a challenge at low currents) the sum of the currents through Q2 and R2 is equal to the current flowing through Q11 to Q18 and R3. Furthermore the operational amplifier regulates the voltage at R2 and R3 to exactly the same value. Thus the currents through R2 and R3 become equal. Since I1 and I2 are equal as well the currents into Q2 and into Q11 to Q18 must be equal too. As a consequence we find the old equation again:

$$V_{R1} = V_T ln(n) \tag{1}$$

So the currents flowing into R1 and into Q2 become:

$$I_{R1} = I_{Q2} = \frac{V_T * ln(n)}{R_1} \tag{2}$$

The parallel path R2 and R3 carries a current of:

$$I_{R2} = I_{R3} = V_{be}/R_2 = V_{be}/R_3 = V_{be}/R \tag{3}$$

assuming R2=R3=R. So the current flowing in M1 and M2 is:

$$I_{M1} = I_{M2} = \frac{V_{be}}{R} + \frac{V_T * ln(n)}{R_1} \tag{4}$$

Now R1, R2 and R3 must be chosen in a way that the current flowing through M1 and M2 becomes constant (except for the temperature coefficient of the resistors). So let us have a look at the derivatives:

$$\frac{dI_{M1}}{dT} = \frac{dI_{M2}}{dT} = \frac{dV_{be}}{dT * R} + \frac{dV_T * ln(n)}{dT * R_1} = 0 \tag{5}$$

In other words:

$$-\frac{R_1}{ln(n)} * \frac{\frac{dV_{be}}{dT}}{\frac{dV_T}{dT}} = R = R_2 = R_3 \tag{6}$$

Still looks terrible? No! We can take it from some lines up:

$$\frac{dV_{be}}{dT} = -2mV/K \tag{7}$$

(what did you expect?) and

$$\frac{dV_T}{dT} = \frac{k}{e} = 86.1733\mu V/K$$

plugging the numbers in we get:

$$R = R_2 = R_3 = R_1 * \frac{23.209}{ln(n)}$$

With this sizing of R2 and R3 the current through M1 and M2 gets constant. So we have the same constant current (or a multiple of a constant current depending on the size of M3) flowing into R4. As a result the voltage at R4 becomes constant. If R4=R3=R2 and M3 is equal M1 and M2 we just will find the bandgap voltage at R4.

1.1 The starter problem of the CMOS bandgap with parallel resistors:

The parallel resistor bandgap suffers from the starter problem. As long as there is no current flow through the bipolar transistors (V(E1) = V(E8) < Vbe) the start up depends on the resistors R2 and R3 only. Ideally R2=R3 and the loop gain is exactly one. So the circuit is weak stable between 0V and Vbe. This means it will neither start nor turn off.

In practice the resistors are slightly different. If R2 is slightly bigger than R3 the operational amplifier will see a slightly higher voltage at the negative input. Provided the current mirror matches and the amplifier has no offset the circuit will start.

The opposite happens if R2 is slightly less than R3. Now the positive input is slightly higher. In the following figure this is the range where the blue line (representing the inverted current I1) is above the red line (representing the inverted current I2). In stead of turning on the amplifier will turn off the current mirror. Thus the bandgap will not start but turn off! Only when the current through the bipolar components increases the voltage at E8 will be lower than the voltage at E1. In the appendant figure this is the range where the blue line is below the red line.

Bottom line we have 3 stable operating points:

1. P1: V(E1)=V(E8R1)=0: The bandgap is off. This point is strong stable if R2<R3.
2. P2: V(E1)=V(E8R1)=V(E8)=Vbe: This is a weak stable point where the amplifier input voltage becomes exactly 0. There is no current flowing through the bipolars yet.
3. P3: V(E1)=V(E8R1) but due to current flowing V(E8R1)>V(E8). This is the desired operating point of the bandgap.

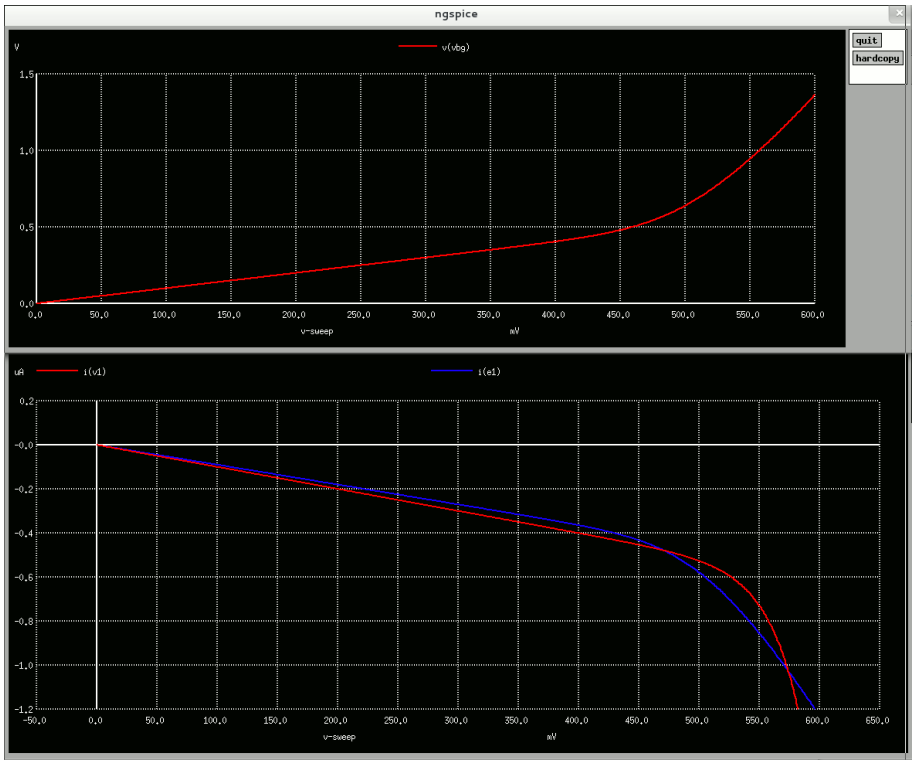

Fig.1.1.1: Start up conditions if R2 is less than R3.

In the above figure the blue line above the red line means we are in a range the bandgap voltage will decrease (move to the left). If the red line is above the blue line the bandgap voltage will increase (move to the right). Looking at the bandgap voltage the weak stable point P2 is found at Vbe. The desired operating point P3 is at bandgap voltage.

For the design of the starter this means the starter may not turn off below P2 (Vbe) but must turn off before reaching P3 (Vbg). Thus the starter MUST correlate with Vbe. In the circuit shown above this means the threshold of M5 must under any operating conditions satisfy the condition:

$$V_{be} < V_{thM5} < V_{bg}$$

Since MOS transistors have a considerable production spread and M5 operates through a wide range of current flowing through starter resistor R5 this simple starter can not be expected to do the job sufficiently reliable! If we have a starter that correlates with Vbe the problem can be solved. But this makes the once very simple bandgap as complex as all the others and the benefit of the simple circuit is lost. (to be precise: due to component leakages P3 may exist in other bandgap topologies too. But there P3 is somewhere around Vout $<$ 10mV and we have no problems with the spread of the threshold of M5.)

1.2 Statistical errors of the CMOS bandgap with parallel resistors:

The bandgap shown above requires the matching of the following components:

1. The bipolar transistors Q2 and Q11 to Q18 must match.
2. The offset of the operational amplifier must be small compared to Vt.
3. The resistors R1, R2, R3 and R4 must match.
4. Current mirrors M1, M2 and M3 must match.

Let us calculate the propagation of errors of these components into the bandgap voltage.

An offset of the bipolar transistors will be multiplied by the ratio of R1 and R4 and multiplication factor of M3. Assuming M3 has the same size as M1 and M2 the multiplication factor becomes 1. Thus we find:

$$\Delta V_{bgbip} = V_{osbip} * \frac{R_4}{R_1} \tag{8}$$

An offset of the operational amplifier propagates the same way (again assuming W/L of M3 is equal to W/L of M1 and M2)

$$\Delta V_{bgOP} = V_{osopamp} * \frac{R_4}{R_1} \tag{9}$$

Resistor spread usually is the bigger the smaller the resistors are. In most designs the area of R1 is the smallest and the spread of R1 is the most significant of all the resistor errors. It propagates as follows:

$$I_{R1} = \frac{V_T * ln(n)}{R_1} \tag{10}$$

$$\frac{dI_{R1}}{dR_1} = -\frac{V_T * ln(n)}{R_1^2} \tag{11}$$

Multiplying this expression with the change of R1 we get:

$$\Delta I_{R1} = -\frac{\Delta R_1}{R_1} * \frac{V_T * ln(n)}{R_1} \tag{12}$$

Now we just have to multiply the change of the current caused by the deviation of R1 with R4 to find the change of the bandgap voltage due to R1 spread.

$$\Delta V_{bgR1} = -\frac{\Delta R_1}{R_1} * \frac{R_4}{R_1} * V_T * ln(n) \tag{13}$$

As long as offsets are caused by area mismatch of bipolar transistors or MOS transistors operating in weak inversion (This usually is the case at well designed operational amplifier input stages) these offsets follow VT. The same applies to the offset caused by a deviation of R1. So these errors can be trimmed without

producing severe problems with temperature coefficients. Looking at the matching of M1 to M3 the situation becomes different. Since we want to have good current matching these transistors usually are designed to operate in strong inversion. We have a parabola shaped characteristic in stead of an exponential one. The offset caused by M1, M2 and M3 mismatch will not follow Vt. The offset errors of the current mirror can only be trimmed out for one temperature but the temperature coefficient remains!

For the calculation of the error contribution of M1, M2, M3 we need to know the operating point of the transistors and the technology properties (gm, tox, matching coefficient..). Empirically the matching coefficients of MOS transistors follow about tox.

$$match \approx \frac{t_{ox}}{nm} * mV * \mu m$$

To calculate the offset voltage we must know the gate area of the MOS transistors.

$$V_{osMOS} = \frac{match}{\sqrt{W * L}}$$

Since we are interested in the current error in stead of the offset voltage we have to figure out the transconductance of the transistors in their operating point.

$$I_d = g_m * \frac{W}{L} * V_{gs}^2 \tag{14}$$

$$\frac{dI_d}{dV_{gs}} = 2 * g_m * \frac{W}{L} * V_{gs} \tag{15}$$

$$\Delta I_d = \frac{dI_d}{dV_{gs}} * V_{osMOS} = 2 * g_m * \frac{W}{L} * V_{gs} * V_{osMOS} \tag{16}$$

$$\frac{\Delta I_d}{I_d} = \frac{V_{osMOS}}{V_{gs}} \tag{17}$$

If the current through M1 deviates from the current through M2 we will see a deviation of the bandgap output voltage of:

$$\Delta V_{bgM1} = R_4 * I_d * \frac{V_{osM1}}{V_{gs}} \tag{18}$$

The same applies if M3 deviates from M1 and M3.

$$\Delta V_{bgM3} = R_4 * I_d * \frac{V_{osM3}}{V_{gs}} \tag{19}$$

Assuming all these errors are statistically independent of each other we have to add the squares and take the square root of them.

$$\Delta V_{bg} = \sqrt{\Delta V_{bgbip}^2 + \Delta V_{bgOP}^2 + \Delta V_{bgR1}^2 + \Delta V_{bgM1}^2 + \Delta V_{bgM3}^2} \tag{20}$$

As mentioned before only the first 3 errors can be trimmed over temperature. So the following is a reasonable worst case estimation for the trimmed bandgap:

$$\Delta V_{bgtrimmed} = \sqrt{\Delta V_{bgM1}^2 + \Delta V_{bgM3}^2} \tag{21}$$

Note: These are one sigma errors! Usually you want to have good production yield and the 1s error should be about 5 to 6 times lower than the specified error (5 sigma design or 6 sigma design strategy).

2 CMOS bandgap with improved accuracy

Th CMOS bandgap shown before has two issues:

1. The poor accuracy due to the high number of error sources.
2. The critical start up circuit.

The following bandgap reduces the number of error sources and solves the start up problem. As a disadvantage it does not provide a reference current for free. It automatically provides a bias current with positive temperature coefficient (ptat current).

$$I_{ptat} \sim V_T/R \tag{22}$$

To create a temperature constant current a current with negative temperature coefficient must be added (ntat current). In the following circuit the ntat current is created using Vbe and a resistor.

$$I_{ntat} \sim V_{be}/R \tag{23}$$

At the end the desired current generator temperature coefficient can be chosen by the ratio of the two currents added.

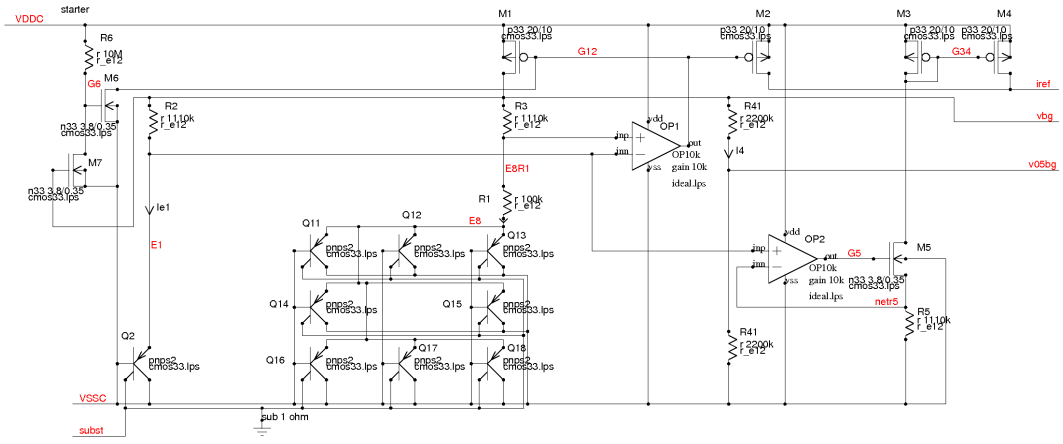

Fig.2.1: Improved CMOS bandgap

Now the bandgap voltage is created inside the regulation loop. The matching errors of the PMOS transistors do not influence the bandgap voltage anymore. The only parameter still affected by the mismatch of the PMOS transistors is the reference current.

The bandgap shown can be regarded as a Brokaw bandgap turned upside down. In stead of using the bipolar transistors as an amplifier stage we now have to use the operational amplifier OP1 because in most CMOS technologies the collector of the PNP transistors is coincident with the substrate. We find the usual equations:

$$V_{R1} = V_T * ln(n) \tag{24}$$

$$V_{R2} = V_{R3} = \frac{R_2}{R_1} * V_T * ln(n) \tag{25}$$

$$V_{bg} = V_{be} + V_{R2} \tag{26}$$

As usual we have to chose the ratio of R1 and R2 in such a way that the temperature coefficients cancel.

$$\frac{R_2}{R_1} * \frac{k}{T} * ln(n) = 2mV/K$$

Propagation of errors to the bandgap voltage: We have 3 main errors affecting the bandgap voltage:

1. The bipolar transistors Q2 and Q11 to Q18 must match.
2. The offset of the operational amplifier must be small compared to Vt.
3. The resistors R1, R2, and R3 must match.

The offset of the bipolar transistors propagates to the output voltage of the bandgap:

$$\Delta V_{bgbip} = V_{osbip} * \frac{R_2}{R_1} \tag{27}$$

An offset of the operational amplifier propagates the same way:

$$\Delta V_{bgOP} = V_{osopamp} * \frac{R_2}{R_1} \tag{28}$$

Resistor spread usually is the bigger the smaller the resistors are. In most designs the area of R1 is the smallest and the spread of R1 is the most significant of all the resistor errors. It propagates as follows:

$$I_{R1} = \frac{V_T * ln(n)}{R_1} \tag{29}$$

$$\frac{dI_{R1}}{dR_1} = -\frac{V_T * ln(n)}{R_1^2} \tag{30}$$

Multiplying this expression with the change of R1 we get:

$$\Delta I_{R1} = -\frac{\Delta R_1}{R_1} * \frac{V_T * ln(n)}{R_1} \tag{31}$$

Now we just have to multiply the change of the current caused by the deviation of R1 with R2 to find the change of the bandgap voltage due to R1 spread.

$$\Delta V_{bgR1} = -\frac{\Delta R_1}{R_1} * \frac{R_2}{R_1} * V_T * ln(n) \tag{32}$$

The total 1s statistical error becomes:

$$\Delta V_{bg} = \sqrt{\Delta V_{bgbip}^2 + \Delta V_{bgOP}^2 + \Delta V_{bgR1}^2} \tag{33}$$

All these errors follow Vt. So trimming the bandgap can be expected to be efficient over the whole temperature range. This is a significant difference to the equations of the bandgap with parallel resistors at the bipolar transistors.

2.1 Propagation of errors to the reference current:

The reference current is affected by the following statistical errors:

1. The error of the ptat current
2. The error of the ntat current
3. the error of current mirror M1, M2
4. The error of the current mirror M3, M4
5. The absolute value of R1 and R5

If the spread of the reference current is a problem depends on the application. The spread of the absolute value of the resistors is in the range of +-20% (at one temperature!). There are two cases to be distinguished:

1. If the reference current is used to supply other cells on the same chip that require a defined voltage drop across resistors of the same type the spread of the resistors absolute value cancels.
2. If the reference current is compared to current flowing outside of the chip the absolute value of the current must be trimmed.

Error of the ptat current:

The ptat current error has 3 main contributors: The offset of the bipolar transistors, the offset of the amplifier and the deviation of R1.

$$\frac{\Delta I_{ptatbip}}{I_{ptat}} = \frac{V_{osbip}}{V_T * ln(n)} \tag{34}$$

$$\frac{\Delta I_{ptatOP}}{I_{ptat}} = \frac{V_{osopamp}}{V_T * ln(n)} \tag{35}$$

The error contribution of R1 is a function of the deviation of R1 due to shape inaccuracies compared to an ideal R1 with perfect shape. $m_R$is the matching number of the resistor R1. It usually is scaled in $\%\mu m$. Typical values for modern technologies are around 3%μm to about 10%μm.

$$\frac{\Delta I_{ptatR1}}{I_{ptat}} = \frac{\Delta R_1}{R_1} = \frac{m_R}{\sqrt{W_{R1} * L_{R1}}} \tag{36}$$

Error of the ntat current:
The most important error of the ntat current generator is the offset of the opamp OP2.

$$\frac{\Delta I_{ntat}}{I_{ntat}} = \frac{V_{osOP2}}{V_{be}} \tag{37}$$

Since Vbe usually is one magnitude bigger than the voltage at R1 the offset requirements of OP2 are much less difficult to be met than the offset requirements of OP1. If OP1 is designed for 0.25mV offset OP2 can typically be designed for 2mV offset spread. Thus the area of the input transistor gates of OP2 can be about a factor 50..100 smaller than the area of the gates of the input transistors of OP1.

Error of the PMOS current mirrors: The same calculation as all current mirrors.

$$V_{osMOS} = \frac{match}{\sqrt{W * L}}$$

"match" is the matching coefficient of the PMOS transistors. Normally this number can be found in the technology documentation. If there is no information to be found a rough guess is:

$$match \approx \frac{t_{ox}}{nm} * mV * \mu m$$

The current error of the mirror is:

$$\frac{\Delta I_d}{I_d} = \frac{V_{osMOS}}{V_{gs}} \tag{38}$$

The gate overdrive Vgs can be calculates as:

$$V_{gs} = \sqrt{\frac{I_d * L}{gm * W}} \tag{39}$$

A lot of equations... Well, probably the most pragmatic idea is to plug it all into a spread sheet. Of course you can just as well design something suitable as

a first guess design and run Monte Carlo simulations until things look nice. But a spread sheet is faster and it pinpoints what is causing most of the trouble much better than a simulation!

And now including the starter: At the end the amplifier OP1 has to be broken down to the transistor level to see if the starter works. Here comes an example:

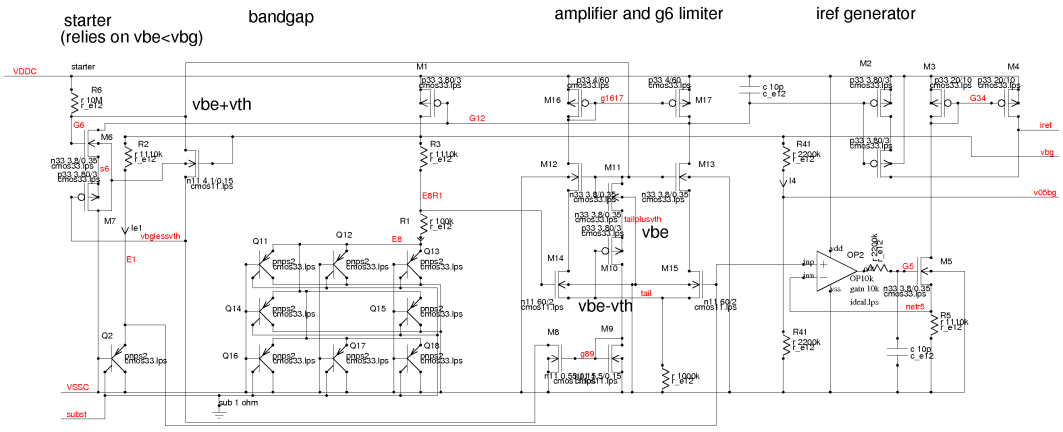

Fig.2.2: The CMOS bandgap including the starter

In the example shown the starter also biases the cascodes. This way in the amplifier transistors with a thinner gate oxide can be used (type n11, M14, M15, M8, M9). This offers a lower offset voltage with less area penalty. (This approach makes sense in modern technologies having single gate oxide transistors for the 1.5V logic and analog circuits and double gate oxide transistors for more rugged I/O cells.)

M8, M9 recycles the current flowing through the clamp M10 and M11. This simply is cheaper than adding an other resistor of several Meg. Ohms.

3 Outlook

In my next post I will have a look at using weak inversion of MOS transistors instead of using bipolar components. This isn't any better, but if your project manager insists on using components that are official parts of the technology and you have no official model of the PNP transistors this may be a way out.