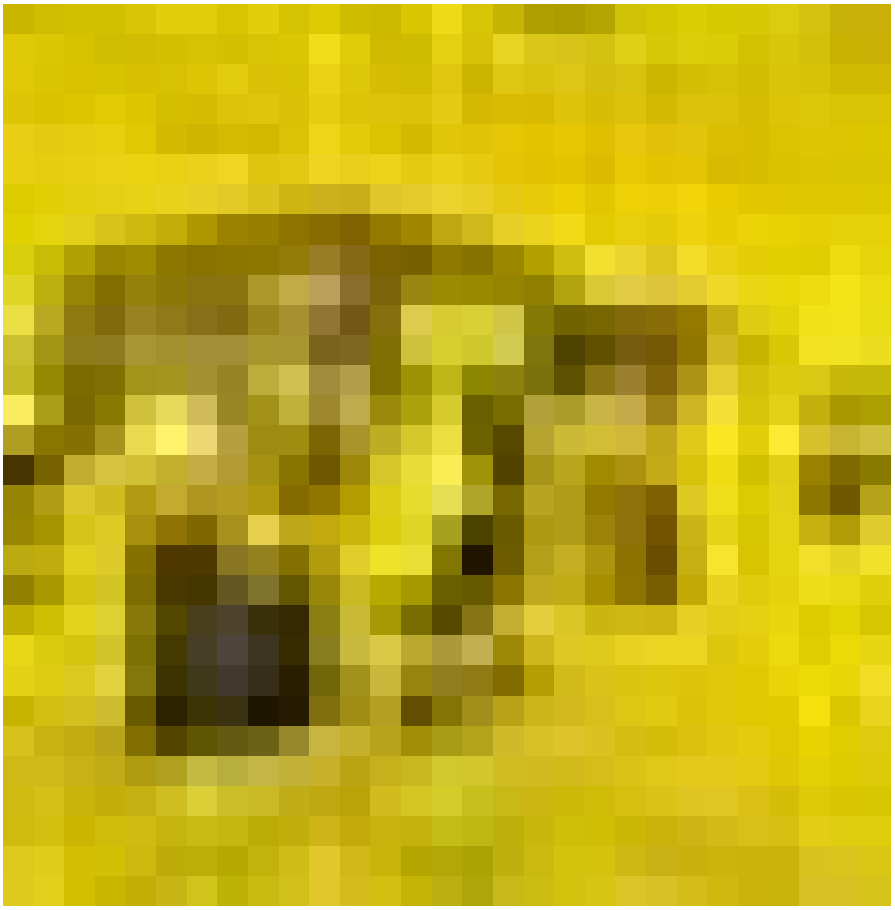
An Overview of BiCMOS Analog IC Design

M501 Term Paper

Stephen H. Lafferty, February, 2007

Contents

Up to 30,000 wafers a month are made at National Semiconductor wafer fab in South Portland. Yellow lighting protects light-sensitive photomask operations.



(decorative)

Photo credits

Front matter: National Semiconductor (wafer-fab), Silvaco (Athena simulation). Conclusions: Texas A&M University (BiCMOS chip).

An Overview of BiCMOS Analog IC Design

Introduction

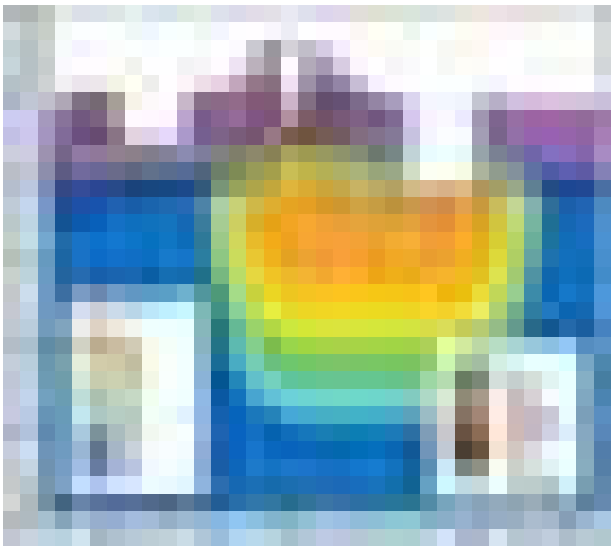
As we approached this brief survey of how bipolar transistors can benefit analog CMOS designs, certain questions seemed to leap out:

- As CMOS is scaled to having supplies of less than 1-Volt, what will happen to the efficacy of the bipolars?
- What performance and cost advantages do bipolars provide? Is that likely to continue?
- What circuit design techniques are used to take special advantage of bipolar devices?
- How has BiCMOS process development fared in recent years? Is it likely to continue being a major part of analog integrated circuit design?
- Why didn't Baker [43] address BiCMOS? Why did Gray's 4th edition [37] cover bipolar and CMOS separately, except for several pages?

In the short time that we had to prepare this paper, we found only one comprehensive, general survey of the BiCMOS landscape. [Late-breaking exception mentioned below.] Fortunately, editor Alvarez' *BiCMOS Technology and Applications, 2nd Edition* [36] is a treasure-trove of information in some of the areas we were researching. The downside is that, since it was published in 1993, BiCMOS has undergone major developments which greatly affected analog design. Harame *et al.* [58] mentions that the SiGe heterojunction bipolar transistor (HBT) made the transition from research to commercialization in 1995. In the communication technology arena, the Joseph et al. paper [28] certainly is a useful and recent survey.

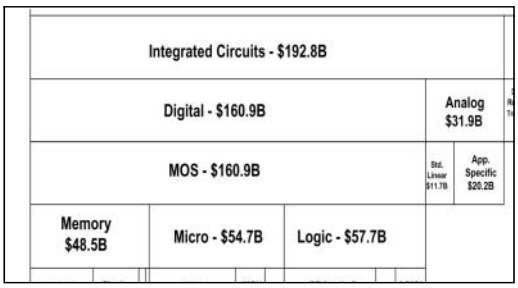
Since the relative merits of bipolar and CMOS transistors take center stage in the discussion of BiCMOS technology, the recent Jagannathan *et al.* paper [29] comparing the two is especially helpful. Many other important references completed our efforts to fill in the time gap since Alvarez. Please note that for the rest of this paper, we will dispense with the unwieldy "*et al.*" notation and refer to papers by the first author only.

Finally, while putting the finishing touches on this paper, we stumbled across an awesome, 38-page survey paper [69] from IBM on BiCMOS technology, processes and applications. It has no fewer than 22 contributing IBM researchers, many of whose names we now recognize. Breathtaking in depth, breadth and clarity, and beautifully published in color, it offers valuable additional insight on many of the topics we cover.



Athena process simulator shows the net dopant profile for a SiGe HBT. Vertical micron scale is 4X magnified over the horizontal. *(decorative)*

Strengths of Bipolar Junction Transistors (BJT's)



The reason that we begin with this topic is that, in browsing the literature, it quickly becomes apparent that the state of the integrated circuit (IC) art is driven mostly by the big *kahuna* of the CMOS digital market (right) [42]. This results in BiCMOS processes used for analog, which are derived from existing CMOS processes. Naturally, the question arises: Why add BJTs? Since this paper is targeted at analog design, we view the BJT's strengths mostly from that perspective.

Noise

Wideband Noise

We know from general experience that MOSFETs can compete well with BJTs on noise at RF frequencies. For many years, the use of MOSFET front ends was *de rigueur* in high quality FM tuners, though low-noise bipolar designs could perform well also. Figure 1 shows noise figure measurements [29] comparing the NMOS and HBT devices selected by Jagannathan. As he noted, the FET seems to do better at low power but can't match the BJT at higher current densities at 25GHz. But there is more to this story. The input impedance for the FET is mostly capacitive, making it hard to match, whereas the BJT source can be closer to 50Ω. The FET noise input resistance is 4-6 times that of the BJT. Thus, the FET can do well but might require off-chip matching inductors. The BJT supports a fully integrated system.

Fig. 1. Noise Figure (F_{min}) vs. current for nFET and HBT at 15, 25GHz.

Flicker (1/f) Noise

For audio-frequency and below applications, the equivalent input voltage of flicker noise of the MOSFET becomes important, and is generally worse than the BJT. Comparing the TLC2201 CMOS opamp to the venerable uA741 in Figure 2 [41], you can see that, while they are comparable at high frequencies, the FET noise voltage is about 3X worse at 10Hz. But even this deficit belies the true magnitude of the problem. Since the TLC2201 is made in TI's LinCMOS™ process we imagine that it represents a much better case than processes stemming from high-performance digital applications. More importantly, TI had to use a tremendous amount of area to achieve this modest capability. In the chip layout shown in Figure 3 [44], the input pins 2 and 3 are at the bottom and the output pin 6 is in the upper right

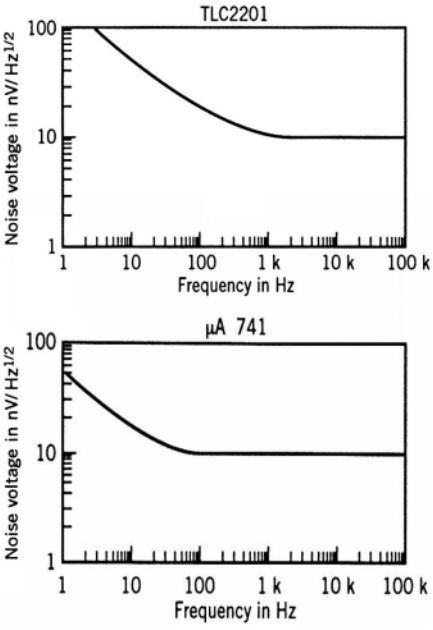


Fig. 2. Input noise voltage for TLC2201 (top) and uA741.

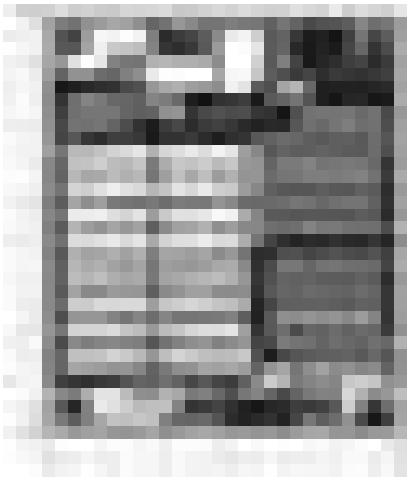


Fig. 3. Die layout of TLC2201.

corner. We surmise that the input FETs are in the large array on the left, and the array on the right contains the output FETs. Thus, input and output are aligned thermally. The input FETs have long gate channels to minimize noise and are interdigitated to minimize offset, while the output FETs have short channels to maximize g_m. We measure the area devoted to the input devices as 34 by 43mils or 860 X 1090μm. Call it 1M μm^2 ! A side-effect of this is a reported 15pF of input capacitance [45], which is not mentioned in [44]. By comparison, BJTs need not be large to exhibit low noise behavior. The HBT used in Figure 1 was just 8 μm^2, about 125,000X smaller. If you insist on using the uA741 from circa 1970 for comparing area, we measure about 8 x 5mils [46] for the input pair or 26K μm^2—still a factor of 38 smaller. Since input noise voltage density (e_n) scales with the root of the gate area, reducing the FETs by that factor would increase e_n by more than 6-times. Combining that factor with the original 3X deficit of Figure 2, we estimate the flicker e_n ratio at 18X for input devices occupying 26K μm^2. By the way, in case you think that the uA741 is too primitive a choice for noise comparison, 1kHz values for the NE5532 are: 5nV/$Hz^{1/2}$, 0.7pA/$Hz^{1/2}$ and for the OP-270: 3nV/$Hz^{1/2}$, 0.6pA/$Hz^{1/2}$. Current noise (i_n) is similar to the uA741. Having the high-frequency e_n comparable to the TLC2201 favors the uA741 for 1/f noise comparison.

To be fair, the MOSFET does have the slight advantage that at low frequencies, i_n is practically zero. For high-impedance sources, the finite current noise of BJTs might be significant. Applications such as photodiode preamps, phono preamps and electret microphone preamps come to mind. To put this in perspective, we can calculate the source impedance at which the current noise becomes equal to the voltage noise. For BJTs, the input noise current generator is the shot noise of the base current [41]. The theoretical noise current is:

$$i_n = \sqrt{2qI_B}$$ in Amps RMS, where q is electron charge and I_B is the base current

So i_n increases with the root of the base current. For the uA741, i_n is about 0.7pA/$Hz^{1/2}$ [41], neglecting flicker noise. Taking e_n at 10nV/$Hz^{1/2}$, the crossover source impedance (Z_X) is about 14kΩ. The NE5532 is at 7kΩ and the OP-270 is at 5kΩ. However, very low e_n makes the effect of i_n and Z_X look bad. For the effect of i_n to exceed the e_n of the FET, source impedance would have to be >15kΩ at 1kHz. The upshot is that, for high source impedances the MOSFET may have an advantage. At low frequencies, the BJT excels.

Offset Voltage

Inherent Performance

The analysis given in [36, pp.308-309] by H.-S. Lee shows that the BJT offset is 5-10 times better than the MOSFET. For the BJT:

$$V_{OS} = (kT/q)\ (-\Delta R/R - \Delta A_E/A_E - \Delta Q_B/Q_B)$$

Where R is the load resistor, A_E emitter area, and Q_B base Gummel number. For the MOSFET:

$$V_{OS} = \Delta V_{TH} + [(V_{GS} - V_{TH})/2]\ [-\Delta R/R - \Delta(W/L)/(W/L)]$$

Lee assumes that the combined Δ-terms are comparable for BJT and MOSFET. He estimates the multiplier term, $(V_{GS} - V_{TH})/2$, at 500mV but to be fair, we update this per Baker [43, p.292] to 250mV. Since kT/q is about 25mV, the BJT is better by a factor of ten, so far. The MOSFET is burdened by an additional ΔV_{TH} term, not found in the BJT, so it gets worse. As Lee concludes, MOSFET diffamps have typical offsets of 5-10mV, compared with 0.5-1mV for BJTs.

Floating Gate Offset Adjustment

The use of nonvolatile floating-gate programming techniques have long been known as a means of trimming the offset voltage of CMOS amplifiers [66]. In that early paper, a typical offset voltage (V_{OFF}) of 10μV was reported, along with drift of 0.7mV over 0-50°C and 0.1mV over 10-years. This level of performance certainly put the MOS transistor offset in the league of the native BJT.

A recent paper from Texas Instruments [67] also uses a floating gate for adjustment and tunneling as an erase mechanism. However, it differs in that hot-electron injection is used for programming. The authors claim faster programming and lower programming voltages for this method. This is implemented in a standard 0.5μm digital CMOS process, with the floating gate structure shown in Figure 4. The reported V_{OFF} is ±25μV, with drift of 40μV over 0-100°C and 0.5μV over 10-years at 55°C. The authors point out that their technique has advantages over alternative trimming methods, including the fact that it can be done at the package level. That eliminates the errors introduced by packaging and contact potentials. Typical programming time is 15s.

Fig. 4. PFET floating gate transistor from [67]. Floating gate node Vfg is surrounded by SiO_2. C_{in} is used in programming by hot-electron injection. C_{tun} is used to erase by tunneling.

We should note that, in a BiCMOS process, the use of floating-gate trimming can be used to trim BJT circuitry too. With the lower initial offsets, this could result in even lower offsets and lower noise.

Bandwidth

The advent of the HBT has given the MOSFET real competition in terms of bandwidth (f_T). As cited by Jagannathan [29], f_T values of 350GHz for the HBT and 243GHz for the MOSFET were reported in the 2003-2004 timeframe. Before trying to compare these devices, we have to decide what criteria to use. Two possibilities are:

- f_T – Frequency for unity current gain.
- G_A – Power gain measured at working bias and frequency conditions, with input matched.

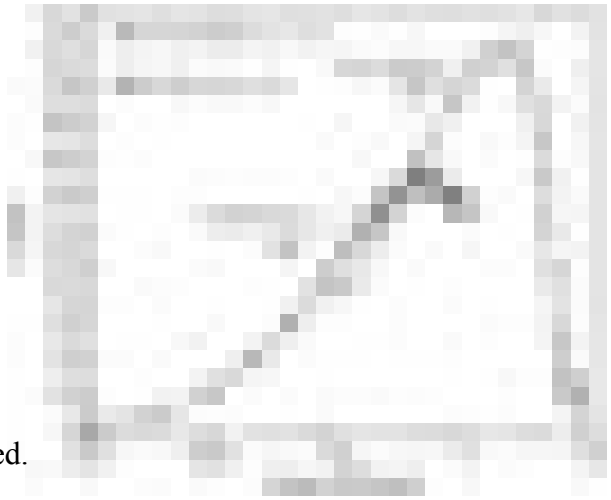


Fig. 5. f_T vs. power comparison of nFET and SiGe HBT.

Jagannathan gives the f_T results shown in Figure 5 (edited here). At low power levels, the FET shows up to a 25%

increase in f_T. This advantage decreases with power and the HBT ultimately yields a 50% increase in f_T over its rival at high power.

Fig. 6. 15GHz Gain (G_A) versus current density for nFET and HBT. The X-axis is expanded on the left.

However, his graph of (power gain) G_A in Figure 6 (edited) seems to tell a different story. Here we are showing the figure in its original scale on the right and a detail on the left, showing the low-current region. The original scale appears to show that the FET can achieve the same maximum gain that the HBT can. In light of Fig. 5, how can this be? The answer is that f_T is based-on current gain which is measured without resonating the FET input capacitance. That sinks most of the input current, reducing current gain. In contrast, G_A is measured with the FET input matched, so the input capacitance does not reduce that gain.

In the detail of Figure 6, the FET achieves a given gain at a much lower current density than the HBT. However, at this point, we must raise a question about whether the choice of FET and HBT devices is really comparable. Unfortunately, we did not find in [29] a discussion of why he thinks these devices are comparable, so let's look at them. From the sizes shown in Figure 6, we surmise that the FET element has a drawn gate length of 0.08μm and a width of 3μm. We take it that 48-elements are combined to make an effective width of 144μm. To estimate FET gate area we have: $A_G = (0.08 \times 3) \times 48 = 11.5\mu m^2$. Since we don't know the area of the source and drain stripes, we cannot really estimate the overall area of the FET, however.

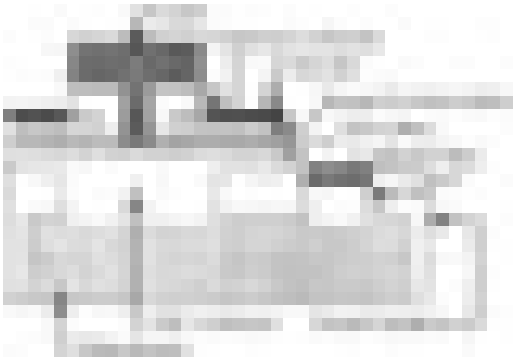


Fig. 7. Structure of the HBT.

To interpret the numbers given for the HBT, we turned to Jagannathan's earlier paper [38], which he referenced as having reported the HBT devices. It indicates that the (0.12×4) term is the emitter size. To estimate the emitter area we have: $A_E = (0.12 \times 4) \times 16 = 7.7\mu m^2$. Figure 7 shows the structure from [38]. Notice that the base and collector regions are rather large compared to the emitter "hat." We can imagine that this area is a bigger factor for the HBT than the source/drain stripes are for the FET and account for the ratio, $A_G/A_E = 1.5$.

To interpret Figure 6, though, we are interested in the linear length of the emitter stripe. That would be: $L_E = 4 \times 16 = 64\mu m$. This is just 44% of the FET length. If we stipulate that Jagannathan's choices of FET and HBT are comparable overall, then a graph based on linear current density gives the FET 2.25-times the total current that the HBT gets. Adjusting for that unbalance would go a long way towards equalizing the gain/power performance in Figure 6, but would still leave the FET with a substantial advantage in low-power operation. Visually, we estimate the current (power) savings as two-times or more at low currents.

Finally, we note that [38] found that HBT f_{MAX} peaked in the range of 1-2mA/μm which is about three times the higher current densities in Figure 6. For a smaller HBT emitter of (0.12×0.5)μm, this resulted in an f_T of 180GHz for 0.8mA current. With ports matched and feedback capacitance neutralized (Mason's gain, **U**), gain was over 25dB at 15GHz. A similar value was reported in [29] for the FET, but the FET is a far larger device.

Summarizing the Bandwidth Comparison
We have found the following points, based-on [29] and [38]:

- In the 90nm process, the HBT can exceed the FET f_T by 1.5-times but this requires high power density.
- At low and matched linear power density, the FET can exceed HBT f_T by 1.25-times.
- Small $0.06\mu m^2$ emitter HBTs, can yield an f_T of 180GHz for 0.8mA current.
- As a result, HBTs can beat FETs in wideband (switching) applications, when such power is not an issue.
- FETs require a far larger area to deliver performance comparable to the smallest HBTs. For example, gate area of the [29] FETs was $11.5\mu m^2$ versus the $0.06\mu m^2$ HBT emitter of [38]. [Note that base and collector probably expand this HBT more than do the source and drain of the FET.]
- In narrowband applications which allow external matching inductors, FETs can equal the high frequency gain of HBTs at substantially reduced power (perhaps 2X or more).
- Due to lower input reactance, HBTs have less need for external input matching networks than do FETs.

Transconductance (g_m)

Benefits of g_m
This parameter indicates the magnitude of the change in collector/drain current for a small change in base/gate voltage. Shortly, we will show that the BJT offers more g_m than does the MOSFET. First, though, let's review why high g_m is important. Some of the benefits of high g_m are:

- Voltage gain is proportional to g_m: $G_V = g_m \times r_o$ where r_o is the stage's output resistance.
- Switching circuit delay is inversely proportional to g_m: $T_d = C/g_m$ where C is the load capacitance [36, p.62].
- Output resistance of an emitter/source follower buffer is inversely related to g_m: $r_{out} = 1/g_m$.
- Input resistance of a common base/gate amplifier is inversely proportional to g_m: $r_{in} = 1/g_m$. Low r_{in} reduces the influence of parasitic capacitance at that node.
- In pole-splitting compensation, the second pole is proportional to g_m of the compensated stage: $\omega_2 \cong g_m C_C/(C_C C_1 + C_1 C_2 + C_C C_2)$ where C_C is the feedback cap, C_1 is the input cap and C_2 is the output cap [43, p.679].
- Output resistance of a cascode amplifier or current mirror is proportional to g_m: $r_{out} \cong g_m r_o^2$ where r_o is the output resistance of the transistor [43, p.638].

BJT Versus FET g_m
Remarkably, the g_m of a bipolar transistor, is the same [37, p.27] for a device of any size, either polarity (npn, pnp) and any material (Si, Ge, SiGe, GaAs):

$$g_m = qI_C/kT \quad \cong \quad 38I_C \text{ at } 300K \quad \cong \quad 38\text{mA/V at } 1\text{mA} \quad [0.76\text{mA/V at } 20\mu\text{A}]$$

…So g_m is directly proportional to current for BJTs.

For long-channel MOSFET processes (say 1μm and above), (nmos) g_m is given [43] by:

$$g_m = \sqrt{2KP\left(\frac{W}{L}\right)I_D}$$ where KP = $\mu_n \cdot C'_{ox}$, μ_n is mobility, C'_{ox} is gate cap/area, W, L are width and length of the device, I_D is drain current.

…So g_m is proportional to the root of current and device width for long channels.

For very short-channel MOSFET processes (say 50nm and below), g_m can become limited by velocity saturation effects [43, p.297] to a value of:

$$g_m = \nu_{sat} \cdot C'_{ox} \cdot W$$ where ν_{sat} is carrier saturation velocity.

…So g_m is proportional to device width for very short channels. Note, however, that ν_{sat} and thus g_m can increase with V_{GS} or V_{DS} somewhat, due to velocity overshoot [43, p.153].

Now let's plug-in some typical values to get a handle on long-channel (1μm) MOSFET g_m:

- Baker [43, p.292] gives KP=120μA/V^2, W=10μm, L=2μm and I_D=20μA for typical analog design with a 1μm process. The scenario is that V_{DD}=5V, V_{THN}=0.8V and V_{GS}=1.05V.
- Since the FET is competing on g_m, let's modify this example to put it in its best light:
  - Use I_D=20μA for a low-current case but also use I_D=1mA.
  - Use minimum L=1μm, allowing matching and g_{DS} to degrade.
  - Considering that BJTs can be tiny (say 7.4μm^2 emitter in a 1μm process), how large is reasonable for the FET? Certainly 10X minimum feature size is a modest gate width. Baker discusses 100X devices in some applications. At 250μm, we are at 10% of a classic analog chip width of 100mils. We will go with that, giving W/L=250. At 250μm^2 gate area versus 7.4μm^2 emitter area for the postulated BJT, this is certainly fair to the FET.

So for the long-channel process (1μm), we have:

$$g_m = \sqrt{2(120\times10^{-6})(250)(20\times10^{-6})} = 1.1mA/V$$ at 20μA

$$g_m = \sqrt{2(120\times10^{-6})(250)(1\times10^{-3})} = 7.7mA/V$$ at 1mA

At 20μA, the BJT has 0.7X the g_m of the long-channel FET. At 1mA, it has 5X the g_m. This performance for the FET is only achieved by using a device more than 33X the size of the postulated BJT (by gate/emitter area). Had they been the same size, the FET would have fared about 6X worse.

Still, it is interesting to note that, at very low currents, it is possible for the long-channel MOSFET to match the g_m of the BJT, if a very large device is used.

For the short-channel MOSFET transconductance, we used a simulation which embodies the typical parameters given by Baker [43, p. 300] for 50nm NMOS:

- V_{THN}=280mV
- t_{ox}=14Å
- V_{DD}=1V

In a typical analog design scenario, he uses:

- W=2.5µm
- L=100nm
- I_D=10µA
- V_{GS}=350mV

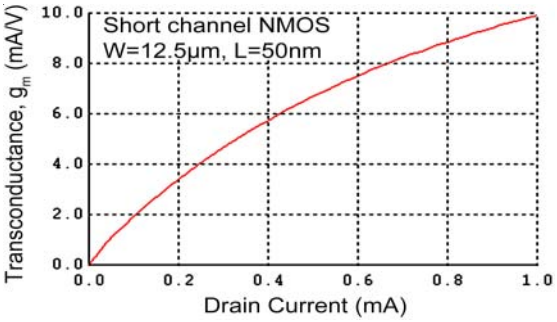


Fig. 8. Short channel NMOS g_m, from SPICE simulation.

To be consistent with our long channel example above, we made these adjustments:

- Use minimum L=50nm.
- Make W=12.5µm, yielding the same W/L as above (250).
- That would scale his 10µA to 100µA but we will sweep from zero to 1mA.
- Bear in mind that in a 50nm process, the BJT emitter size would be smaller than 0.12µm by 0.5µm, which is the size used in the [38] 90nm process. Charitably taking this to be our comparable BJT, we get the gate/emitter area ratio as: (0.05×12.5)/(0.12×0.5) = 10.4.

We ran the SPICE simulation of Figure 8 using Baker's NMOS short channel model [47], which is reproduced in Appendix A. It is interesting that at 1mA, we have g_m=10mA/V, about 1.3X the long channel device of the same W/L. Though it's hard to see on this scale, another run yielded g_m=0.475mA/V at 20µA, about 0.43X the long-channel case. Keep in mind the fact that short-channel FETs are not expected to follow the classic long-channel calculations.

For comparison, let's look at Baker's g_m figures for long and short channel. He gives:

- Long channel:
  - W/L = 5
  - I_D = 20µA
  - g_m = 150µA/V
- Short channel:
  - W/L = 25
  - I_D = 10µA
  - g_m = 150µA/V

If we naively scale the long channel g_m to the short channel case, the "expected" g_m = 237µA/V. The reported g_m = 150µA/V, which is about 0.63X "expected." This is in the ballpark of the distorted expectations, above. We conclude that the results in Figure 8 are not unreasonable, in this light. We also note that, in spite of 14X thinner oxide, the short channel devices do not appear much stronger, in terms of DC transconductance, in these conditions. Superior AC performance comes from smaller area, reduced by the square of the scale factor.

Going back to our BJT-MOSFET comparison: At 20µA, the BJT has 1.6X the g_m of the short-channel FET. At 1mA, it has 3.8X the g_m. This performance for the FET is only achieved at using a device more than 10X the size of the postulated BJT (by gate/emitter area). Had they been the same size, the FET could have fared about 3X worse (using the long-channel equation).

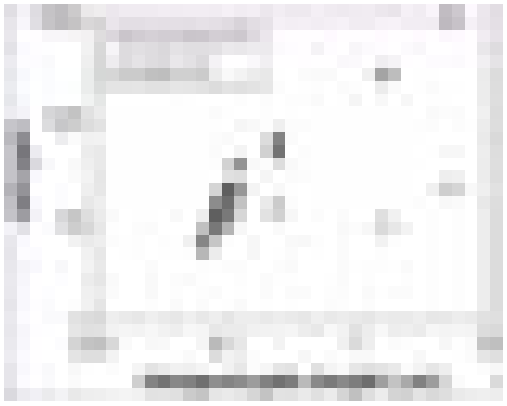


Fig. 9. Stage gain versus gate length in 90-nm CMOS. Measured at 10X the threshold current.

Voltage Gain

DC voltage gain (A_V) is, of course, important in feedback amplifiers, since it helps determine the accuracy of the output signal. A general expression for no-load $A_V = g_m \times r_o$, where r_o is the output resistance of the stage. Since g_m is considered separately above, that leaves us with r_o as the remaining factor in A_V. Thus, we often find concern in the literature about the r_o of transistors and subcircuits. Unfortunately, as CMOS processes are scaled, r_o becomes lower due to increased channel length modulation as the channel length gets shorter in relation to the drain depletion width [43, pp.144, 298].

Figure 9 [28, p.1544] shows that the issue is significant for 90nm CMOS. While DC gain performance has classically been an issue for low-frequency amplifiers, when the stage gain deteriorates into the ~20dB range, it can become a limiting factor for RF performance as well. That is approaching the typical gain expected from an RF stage.

As explained by Jagannathan [29, p.117], one can alleviate this problem by employing a cascode circuit. However, unless the typical ~1V power supply voltage is increased, this can reduce f_T a substantial amount. Increasing voltage to avoid this, results in an increase in power consumption. Power is a key tradeoff factor with bandwidth, as discussed above.

In contrast with the effect of process development on FET gain, the advent of the HBT tended to increase self-gain of the BJT. This is due to the increased base doping [14, slide 18], which reduces base width modulation (BWM). For collector voltages well below breakdown, BWM is the main factor in BJT output conductance [37, pp.13-14]. However, as noted in [48, p.18], with the low breakdown voltage of modern BiCMOS HBTs, one has to be careful about nonlinear output conductance at higher currents.

Below the breakdown region (and above the collector saturation voltage), output conductance of BJTs is modeled in terms of Early Voltage (V_A) as shown in Figure 10 [49]. The output resistance of the transistor [37, p.30] is given by:

$$r_o = \frac{V_A}{I_C}$$

To get a handle on r_o, we scanned the literature to find typical values of V_A. The results of this quick survey are in Table 1. It should be noted that V_A has an inverse relationship with temperature [50, p.28].

Fig. 10. Approximate definition of Early Voltage for a BJT.

Table 1. Early voltages and other parameters reported for BiCMOS HBTs.

Process	V_A	Emitter Size	f_T	Scale	BV_{CEO}	Ref
IBM 5AM HP	48V	0.5 x 1.0μm	50GHz	0.5μm	3.3V	[50]
IMEC 0.25 CMOS	>50	0.25 x 1.65	95	0.25	3.55	[51]
"1st Generation"	65	0.42 x ?	47	0.42	3.3	[52]
"2nd Generation"	120	0.18 x ?	120	0.18	2.5	[52]
"3rd Generation"	>150	0.12 x ?	207	0.12	1.7	[52]
IBM 9T (4th Gen)	>150	0.12 x ?	285	0.12	1.7	[50]
SiGe LN2	113	0.5 x 2.5	57	0.5?	2.7	[52]

From these figures, there seems to be a trend towards higher V_A values in later generation processes. We take the value of V_A=110V as a good conservative typical for a modern process. Since $g_m = qI_C/kT$ and $A_V = g_m \times r_o$, we can combine these with the equation above to get:

$$G_V = \frac{qI_C}{kT} \times \frac{V_A}{I_C} = \frac{qV_A}{kT} = \frac{V_A}{26mV} \qquad \text{at 300K}$$

So for our nominal value of V_A, we have a stage gain of 4200 or 72dB, which is stellar! As noted above, the MOSFET's stage gain of about 10 requires circuit measures to make it usable.

Resistance to Electrostatic Discharge (ESD) Damage

As is well known, CMOS devices are subject to damage from static voltages, due to the fragility of the gate oxide. Although protection circuits are used at input/output pads, the IC is still subject to damage if careful precautions are not taken. For IC's intended to act as high-speed interfaces to external equipment, system designers must add protection networks. To some extent, these networks may impair the signal path. This is particularly true in such wideband interfaces as HDMI, which have tight requirements on capacitive parasitics between the chip and connector.

With gate oxide thicknesses expected to drop to 20Å and below, gate voltage must not exceed 2V or so, or damage will occur [43, p.154]. Bipolar devices, on the other hand are relatively robust with regard to static damage, since reverse breakdown of diode junctions is not necessarily catastrophic. Rather, the junction conducts current and can dissipate the charge safely [53]. Of course, there are limitations to this robustness. With extremely small device dimensions, a static surge of sufficient magnitude can still deliver enough energy to damage a BJT. For devices interfacing with the external world, though, designers can size horizontal dimensions to accept greater stress, subject to performance constraints. As a result, the HBT should offer much better reliability for external interfacing than does the MOSFET.

Summary of BJT Strengths

The following table shows that the BJT has much to offer in BiCMOS analog designs:

Table 2. Summary of BJT Strengths

Parameter	BJT (HBT)	MOSFET (NMOS)
Wideband Noise	Excellent.	Comparable but may require external inductor matching.
1/f Noise	Excellent regardless of size. Advantage decreases for Rsource >>14K.	Reduction requires large device. Still 18X worse at 26K μm^2.
Offset voltage	At least 10X better. Typically 0.5-1mV.	Typically 5-10mV. (See Note-3.)
Bandwidth	Excellent at small size (0.06μm^2). 1.5X better f_T for digital but needs ~1mA. Less need for matching inductors.	Requires larger device (11.5μm^2). With external matching, can equal RF gain at half the power.
Transconductance	Excellent at small size (0.06μm^2). 38mA/V at 1mA. Proportional to I_C.	Requires 10X larger device to match HBT at ~10μA. Still has 4X less at 1mA.
Voltage gain	~4200.	~10. Cascode can help but reduces f_T and increases power.
ESD strength	Larger devices are relatively robust.	Very fragile.

Notes:
1. Device areas are based-on emitter and gate dimensions. Total areas may be different.
2. Comments are intended as trends and typical examples rather than absolute fact.
3. Floating gate adjustment techniques can drastically reduce offset voltage.

All-CMOS Counterexample

In the last section, we made the case which motivates the addition of HBT devices to CMOS processes. As noted though, there are often remedies which allow all-CMOS designs to deliver impressive performance in applications in which HBTs are strong. To put this in perspective, we offer this example.

106dB SNR Sigma-Delta ADC

This chip is shown in Figure 11. Presented by Analog Devices in 2005 [25], it is remarkable due to its low-noise at audio frequencies and its linearity. Implicit in the low-noise sampling is the requirement for very low

Fig. 11. Analog Devices 106dB sigma-delta converter.

jitter. The engineers have used clever design techniques to compensate for the MOS noise and gain issues which would otherwise limit performance in these areas. This is achieved with low operating power and chip area. Key specifications include:

- SNR (A-wtd) 106dB
- THD+N 95dB
- Power 32mW/chan
- Area 1.82mm^2
- Sample rates 8 to 48kHz

Fig. 12. Loop filter block diagram.

How They Did It

The design was implemented in a 0.35µm, 3.3V CMOS process. Some key features we noticed in the design are:

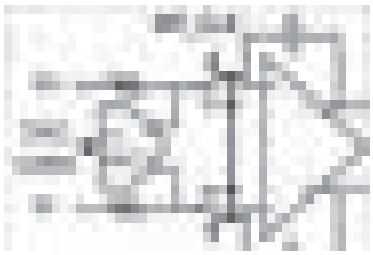


Fig. 13. Input stage.

- The SNR is A-weighted. While not a circuit design technique, it is an important part of the design strategy. Since A-weighting recognizes the value of concentrating on higher frequency noise in audio applications, it preserves the significance of the noise specification while alleviating the major issue of 1/f-noise in MOS transistors.
- A combined continuous-time and discrete-time loop filter structure (Figure 12), with a patent-pending ISI-zeroing scheme sums **ΔΣ** feedback directly at the input pins (Figure 13). This eliminates the need for an input buffer and suppresses signal-dependent digital noise, both of which would limit linearity. It also keeps opamp speed requirements and influence low.
- Full differential design with common mode servo feedback to the input pins also reduces signal-dependent digital noise.
- The resistive summing-junction differential input offers two strengths:
    - Since the differential input gets zeroed each cycle, the designers can use that opportunity to zero the opamp offset and *noise fluctuations* (speculation). That reduces opamp noise drastically in the audio range, since the clock is 1-6MHz.
    - Full rail-to-rail differential input swing is supported. The resulting 2VRMS differential swing maximizes SNR in the 3.3V process.
- A dynamic input range scaling technique involving feedforward noise shaping reduces the size of the integrating capacitance.
- Low jitter (32ps) timing control loop cleans-up jitter on the system clock. Uses same low-noise voltage reference as the DAC in its low-noise relaxation oscillator.

Importance of the SiGe HBT

We began the research for this paper with Alvarez' anthology on BiCMOS, published in 1993. Although the SiGe heterojunction bipolar transistor (HBT) was a topic of research, it was not featured there, since it went from lab to production in 1995. In our literature search to update the book, we were struck by the dearth of information on plain silicon BiCMOS. In its place, the HBT was pretty much the only thing to be found. Initially unfamiliar with the device, we feared that it would require lots of study. In spite of the imposing name though, it generally requires only minor changes at the circuit design level and is very compatible with standard CMOS processes at the physical level. It mainly increases f_T a factor of two or more, over its silicon cousin [37, p.153].

With a non-self-aligned emitter, the number of masks can be kept similar to a conventional BiCMOS process [12]. The recently reported process from Jazz offers 90GHz HBTs, 0.13µm CMOS with a mask-count resembling RF CMOS [9].

Structure of the HBT

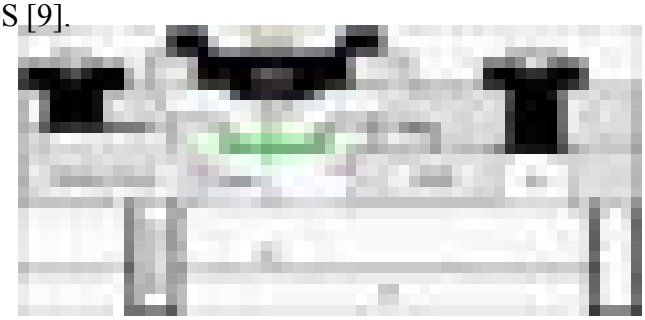


Fig. 14. SiGe HBT. [52, slide 12]

The device gets its name, "heterojunction," from the fact that it uses dissimilar materials on either side of the semiconductor junction. For commercial BiCMOS processes, a SiGe alloy is deposited to form the base, perhaps 25% Ge [50, p.8]. As shown in Figure 14, the collector is silicon and the emitter is polysilicon. An important feature enhancing performance is graded base. That technique varies the Ge content of the base from say, 3% at the emitter to perhaps 25% at the collector. This results in an electric field which accelerates carriers across the base, reducing transit time and increasing f_T [52, slide 20].

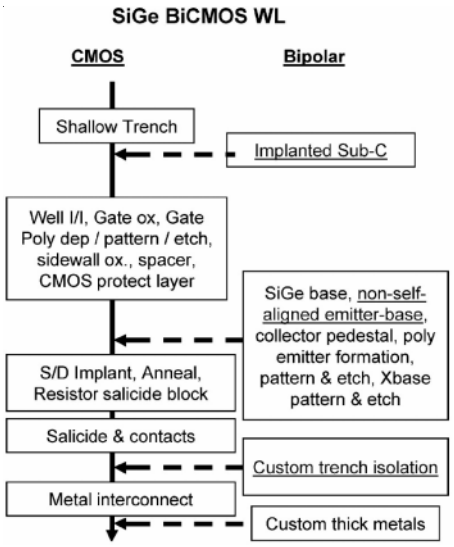


Fig. 15. Integrating HBTs into CMOS process flow.

Although the HBT structure looks complex, the fabrication steps are readily integrated into standard CMOS processes [28, p.1539]. Figure 15 shows the integration of HBT processing into the CMOS flow [28]. The "WL" type process flow is a lower cost approach, targeted at consumer applications. High performance "HP" processes use a self-aligned emitter-base. While the scope of this paper does not include detailed coverage of BiCMOS processes, we have included an illustration of the process flow in Appendix B.

IBM was the first company to commercialize HBT BiCMOS [50] and is now on its fourth generation (120nm). They offer the HBT in two versions: high performance (HP) and high BV_{CEO} (HB). Confusingly, both of these devices are available in both WL and HP *processes*. In the HP process, the HP HBT f_T is 285GHz and BV_{CEO} is 1.77V. The HB HBT is >50GHz and 3.6V.

Table 3. Silicon BJT and SiGe HBT Performance

Parameter	Silicon BJT	SiGe HBT
Beta	67	261
f_T	38GHz	57GHz
BV_{CEO}	3.5V	2.7V
Noise Figure (2GHz)	0.6dB	0.28dB
1/f Noise (Note-1)	$2x10^{-22}$ A^2/Hz	$1x10^{-23}$ A^2/Hz
Early Voltage	19V	113V
Low temp (Beta, f_T, NF at 77K)	All degrade.	All improve.

Note-1. Here, 1/f is S_{IB} at 100Hz. S_{IB} is base current noise density in A^2/Hz. This is found to be the dominant 1/f noise source in BJTs [54].

HBT Advantages Over Conventional BJTs

Table 3 shows a comparison between the conventional silicon BJT and HBT performance [52, slides 22-24]. The devices are not state-of-the-art, they are a reasonable basis for comparison.

Considering that the HBT uses much the same fabrication process as does the BJT, it is a "no-brainer" to use it in preference. The only parameter which *doesn't* improve is BV_{CEO}. Since the supply voltage of the corresponding generation of CMOS is below this, it is not generally an issue. As we saw above, some foundries find it easy to include higher voltage HBTs as well.

HBT Market Position

We have already shown in our first section how the HBT's strengths give it certain technical advantages over the MOSFET. Now we add applications and cost to the picture to show its importance to the market. Figure 16 illustrates the relationship of transistor f_T to market applications for successive generations of process [adapted from 28]. From this we see that the HBT has continued to have advantages over the MOSFET. Moreover, it is becoming a key enabler for emerging products in the 60GHz range. The f_T advantage, perhaps a factor of two, proves decisive for ambitious, cutting-edge applications. This also brings to mind the microprocessor wars and a comment in Alvarez [36, p.380]. Writing as far back as 1993, he noted that Intel's Pentium microprocessor was using BiCMOS to advantage. (Not even HBT!) With Intel and AMD now playing leapfrog in terms of CPU speed, an advantage in the 50% range seems important to the market.

Fig. 16. Effect of f_T on market applications over process generations.

Notice that the HBT f_T improvement over the BJT, roughly accounts for the improvement over the MOSFET. Thus, the advent of the HBT saved the BJT from being eclipsed, explaining the ubiquity of the HBT in BiCMOS processes today.

Intel and BiCMOS

Intel used BiCMOS for at least three generations of the Pentium Processor, at the 0.8, 0.5 and 0.35um nodes. [16]. They apparently abandoned the BJT for new CPUs after 1996 [35], though they do use SiGe strained technology in CMOS processes. Shipments of 45nm CMOS Penryn CPUs are scheduled to begin by the end of 2007.

HBT Characteristics

The subject of BJT devices is something of a chestnut, so we will only discuss highlights and HBT oddities here. See, for example [37], from which we extract some of the following. Some typical HBT parameters were listed earlier in Table 1.

Low Frequency Parameters

First, as noted earlier, g_m is the same for *all* BJTs and is proportional to current. Recall also that dynamic emitter resistance is $1/g_m$ and that it is the emitter follower's output impedance:

$$g_m = qI_C/kT \quad \cong \quad 38I_C \text{ at } 300K \quad \cong \quad 38\text{mA/V at } 1\text{mA} \quad [0.76\text{mA/V at } 20\mu\text{A}]$$

Also mentioned earlier is the output resistance:

$$r_o = V_A/I_C$$

The base-emitter voltage, V_{BE} is:

$$V_{BE} = V_T \times \ln\left(\frac{I_C}{I_S}\right) \qquad \text{where:} \qquad I_S = \frac{qA\overline{D}_n n_i^2}{Q_B}$$

A = emitter area
n_i^2 = intrinsic carrier concentration
$\overline{D}_n$ = avg base diffusion constant
Q_B = base doping per emitter area

Notice that I_S drops with an increase in base doping, increasing V_{BE}, for a given I_C. For a conventional BJT we can calculate the current gain as:

$$\beta_F = \frac{1}{\frac{W_B^2}{2\tau_b D_n} + \frac{D_p}{D_n}\frac{W_B}{L_p}\frac{N_A}{N_D}}$$

W_B = width of base
τ_b = base minority carrier lifetime
D_n = diffusion constant for electrons
D_p = diffusion constant for holes
L_p = diffusion length for holes in the emitter
N_D = emitter doping
N_A = base doping

For high frequency BJTs, it's a given that the base (W_B) is thin. Thus, we get high current gain in the conventional BJT by using a high emitter-to-base doping ratio (N_A/N_D). That tends to push for low doping in the base, since the emitter gets maximum doping anyway. Lower base doping makes the B-C depletion layer larger, reducing V_A and BV_{CEO}.

For the SiGe HBT, we change the base from pure silicon (Si) to an alloy of Si and germanium (Ge). There is a difference in bandgap energy for Si and Ge, so the alloy base has a lower bandgap than pure Si. The effect of this change in bandgap is to increase the flow of collector current, thus increasing beta. The beta enhancement factor of SiGe can be approximated by [50]:

$$\frac{\beta_{SiGe}}{\beta_{Si}} \approx \frac{\tilde{\gamma}\tilde{\eta}\Delta E_{gB}}{kT} e^{\left(\Delta E_{g0}/kT\right)}$$

$\tilde{\gamma}$ = density of states ratio of Si and SiGe < 1.
$\tilde{\eta}$ = electron diffusivity ratio between SiGe and Si > 1.
ΔE_{gB} = change in bandgap across the base due to Ge grading
ΔE_{g0} = change in bandgap from emitter to base

The key to the HBT enhancement is the effect of the change in bandgap due to Ge in the base. Ahmed [50] estimates this as -7.5meV per 1% of Ge. At 300K, kT=26meV. For a Ge grading of 3% at the emitter to 23% at the collector, the most controllable part of the beta enhancement expression becomes:

$$R_\beta = \frac{\Delta E_{gB}}{kT} e^{\left(\Delta E_{g0}/kT\right)} = \frac{(7.5meV \times 20)}{26meV} e^{\left((7.5meV \times 3)/26meV\right)} = 13.7$$

So, ignoring the secondary factors, the SiGe base greatly boosts the beta and hence f_T as well. But not all of this factor is used for gain enhancement. Having this extra degree of freedom lets us trade some gain for higher base doping. That increases Early voltage and lets us make the base thinner for a given BV_{CEO}. The thinner base directly increases f_T. Further, grading the Ge concentration imposes an electric field which sweeps carriers across the base, reducing transit time and further increasing f_T.

High Values of V_{BE}

One factor that we noticed about some reported HBTs is that their high frequency performance seems to peak at high values of current density and V_{BE}. For example, the peak f_T shown in Figure 5 occurs for V_{BE}=0.88V. In another case, measurements and simulations of HBTs featured in a Silvaco brochure show f_T peaking above 0.9V of V_{BE}. As we noted above, increasing the base doping leads to this. Since some CMOS operating voltages are down to the 1.2V range already, this may push BiCMOS designs to require more than one supply voltage.

Low Values of V_{CE}

Stemming from a very thin base, low HBT BV_{CEO} dovetails well with reduced supply voltage. As seen in Table 1, the trend is to reduce breakdown voltage with each generation, with 1.7V being the current value. Collector and V_{BE} characteristics for IBM's first-generation HBT are shown in Figure 17 [50]. This was a 3.3V process with a peak f_T of 47GHz. Notice the V_{BE}'s running around 0.95V. The base currents run 7-13μA, indicating β=100 at the center sweep.

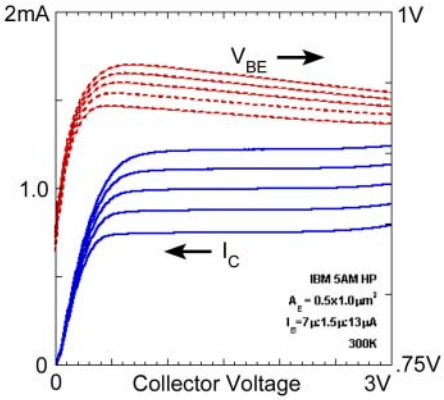


Fig. 17. Collector and V_{BE} data for IBM's first gen HBT.

What we find, then, is that the HBT can be treated just like a conventional BJT for design purposes. It has the same g_m as other BJTs. We should keep in mind, the high f_T, high V_A, somewhat high V_{BE} and low BV_{CEO}.

BJT Transition Frequency (f_T)

We will use the equivalent circuit of Figure 18 to write an approximate expression for f_T of the BJT [37]. The base is driven with an AC current source and the AC collector current, i_o, is measured. The frequency at which the current gain asymptotically becomes unity, is f_T. In practice, this is measured at frequencies well below f_T.

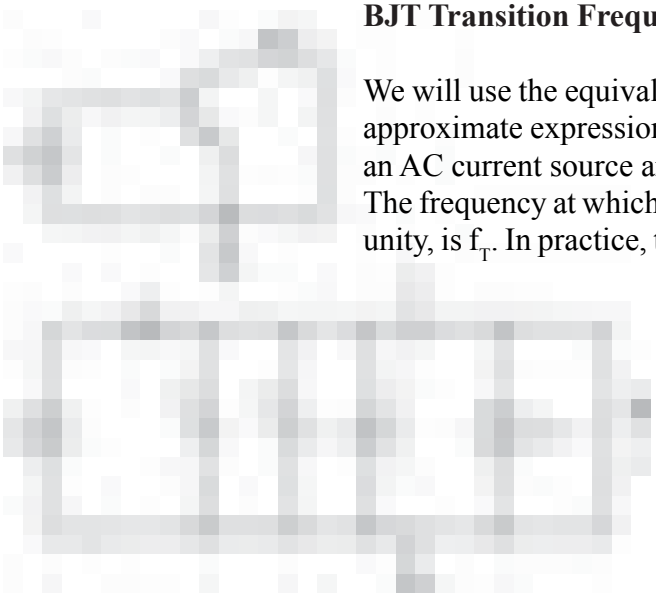


Fig. 18. Simplified small-signal equivalent circuit for calculating f_T.

f_T is dominated by the base charging capacitance, C_b, and the B-E junction capacitance, C_{je}. Further, unless we are operating at low collector current, $C_b \rangle\rangle C_{je}$.

C_b arises from the storage of carriers in the base, due to the base transit time, τ_F. This can be calculated by:

$$\tau_F = \frac{W_B^2}{2D_n}$$

As you would expect, transit time increases with base-width. Beware, though, that actual values

for graded bases may vary from this estimate. Next, we can derive a value for C_b from:

$$C_b = \tau_F g_m = \frac{\tau_F I_C}{V_T}$$ ...In effect, τ_F is the time constant of C_b and r_e (which is $1/g_m$).

We can calculate f_T using this. First define $\omega_T = 2\pi f_T$ and $\tau_T = 1/\omega_T$. We can write:

$$\tau_T \cong r_e(C_b + C_{je})$$ where $C_\mu << (C_b + C_{je})$.

If we are not at low I_C, we can further simplify this to:

$$\tau_T \cong r_e C_b$$

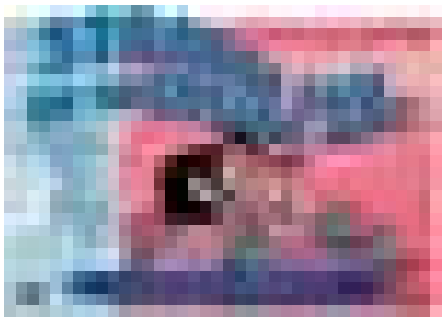
Now, substituting the expressions for C_b and r_e, as well as working back to f_T, we have:

$$f_T = \frac{1}{2\pi\tau_F}$$ or, at low current: $$f_T = \frac{1}{2\pi(\tau_F + r_e C_{je})}$$

Falloff of f_T at High and Low Collector Current

At low values of I_C, f_T falls off because r_e becomes large, increasing the effect of C_{je}. You can see this in Figure 5. At higher currents, τ_F dominates and τ_T approaches τ_F. In other words, peak f_T is the reciprocal of base transit time (with radian-Hz scaling). As I_C continues to increase, f_T eventually peaks and starts to fall off, often very rapidly as is also seen in Figure 5. Two reasons for this are the Kirk effect and reduction of β at high injection. The reduction of β can occur because at high I_C the carrier concentration injected into the base may exceed the doping concentration. The result is that further increase of I_C is diminished [56]. In this region, β is inversely proportional to I_C [37, p.25]. The Kirk effect occurs when the minority carrier concentration in the collector approaches the doping level there. The base of the transistor then enlarges into the collector [37, p.26]. Since the base is extremely thin and its thickness has a dramatic effect on τ_F (hence f_T), one can easily imagine that this accounts for the rapid rolloff seen in Figure 5. Finally, it should be noted that the hybrid-π model, from which Figure 18 and these calculations are derived, is only valid for $f << 3f_T$ [55, p.355].

The Strange Case of the HBT Emitter



Late in the preparation of this paper, we came upon something of a mystery about the HBT. Investigating this strangely hidden issue reminds us of a classic Bell Telephone educational film shown to schoolchildren in the 60's: *The Strange Case of the Cosmic Rays* [60]. We call the emitter mystery strangely hidden, because it emerged only after digging through over 60 references related to BiCMOS; yet it potentially has a significant impact on device and circuit design.

Discovery of the Hidden Layer

Our story begins with your author casually browsing through the BiCMOS research papers. While reading Markus' 1995 "Low-Frequency Noise in Polysilicon Emitter Bipolar Transistors," I was intrigued by his conclusion that the 1/f noise is due to *oxide between the polysilicon (poly) and silicon portions of the "emitter."* I was surprised that there is oxide there. How does the current get across it?

Secret Tunnel

Digging in Teplik's chapter in Alvarez [36, p.37] yielded a cryptic comment about a tradeoff between beta and R_E, depending on preclean oxidation and something about a recombination velocity at the silicon/poly interface. "Tradeoff" implied that it was intentional and controlled. Why hadn't I heard more about this, especially if it affects beta? A search in Teplik's references, revealed de Graaff's 1978 "SIS Tunnel Emitter" paper [61]. It mentions that poly emitter transistors with a 20Å oxide separator have higher beta. Also mentioned is the fact that such beta has a low positive or even negative coefficient of temperature. What does that mean for Lee and Kim's [57] BGR curvature compensation, using base current? (Discussed later.) De Graaff uses a tunneling model to answer the question of how the current gets across the oxide layer. Current through the 20Å layer may be promoted by band-bending at the layer. This is ascribed to a combination of a pileup of dopant and positive charges in the layer.

Invisible Barrier, Missing Evidence

De Graff went on to conclude that the oxide has a voltage barrier up to ~100mV. (What does that mean for the BGR and for V_{BE}-related design equations?) It was unfortunate that IEEE Explore didn't show forward citations for this paper, so I went looking in Harame's comprehensive tutorial on SiGe HBT technology [58]. Strangely, there was no mention of this oxide layer, even though the title purports to treat, "Materials, Physics, and Circuits." Harame's Part II article [59] does cover 1/f noise, noting that the ~300Hz corner is typical for BJTs, but again doesn't mention an oxide sandwich in the emitter. The Vempati *et al.* 1996 paper [62] on LF noise in HBTs *does* mention "interfacial oxide" and notes that they agree with Markus that it is the cause of 1/f.

Are We Chasing a Scary Ghost?

At this point, we began to wonder if the reason that the layer isn't more prominent could be that it doesn't have much effect. The numbers in [61] were scary but his measurements dealt with relatively thick oxide (~20-60Å). Researching the issue further produced [63] which assured us (1) that the thickness of the layer can be well-controlled and (2) that the resulting emitter resistances are low, at least in the size devices they were using. There was no mention of a barrier voltage. He also led us to the Ning-Tang intercept method of R_E measurement [65] which made it clear that they were not expecting any deviation from typical BJT Gummel plot behavior for the poly emitter devices. This was reinforced by [64], which reassured that:

- The oxide effects are accounted-for by a resistor model.
- The emitter resistance can be made minimal.
- Oxide thickness can be kept down in the 8-20Å range
- (Most important?) The oxide is often broken-up, leaving some direct poly/Si contact.

Mystery Solved

Finally, there was the realization that, even if the oxide layer can cause a small, fixed, voltage offset, it would add to V_{BE} and not distort the shape of the exponential current relationship. In effect, it would factor the diode junction saturation current, I_S. However, the fact is that the oxide layer can be broken up, allowing direct poly-monocrystalline silicon contact. That should effectively eliminate the offset. In conclusion, all of these papers treat the oxide-layered HBT as an ordinary BJT (with some R_E) and there is an explanation for how this is possible. Thus, there is no reason think that the oxide layer invalidates treating the HBT as a normal BJT. This explains why the oxide layer is not a more prominent issue in the literature.

BiCMOS Circuit Design Examples

Since this paper is focused-on BiCMOS Analog IC Design, it's about time that we get into some circuit design. Classic bipolar analog design is well covered in the literature [37], so we will emphasize techniques which exploit the integration of MOSFET and bipolar transistors. Also, our generally limited scope only allows superficial browsing of these designs.

Opamp with Common Mode Feedback (CMFB)

The opamp shown in Figure 19 is used in the sample-and-hold of a 10-bit, 44MS/s ADC reported from Texas A&M University [1]. This system provides two ADCs to capture I&Q baseband channels for a multipurpose receiver. The applications are for 11Mb/s 802.11b and 1Mb/s

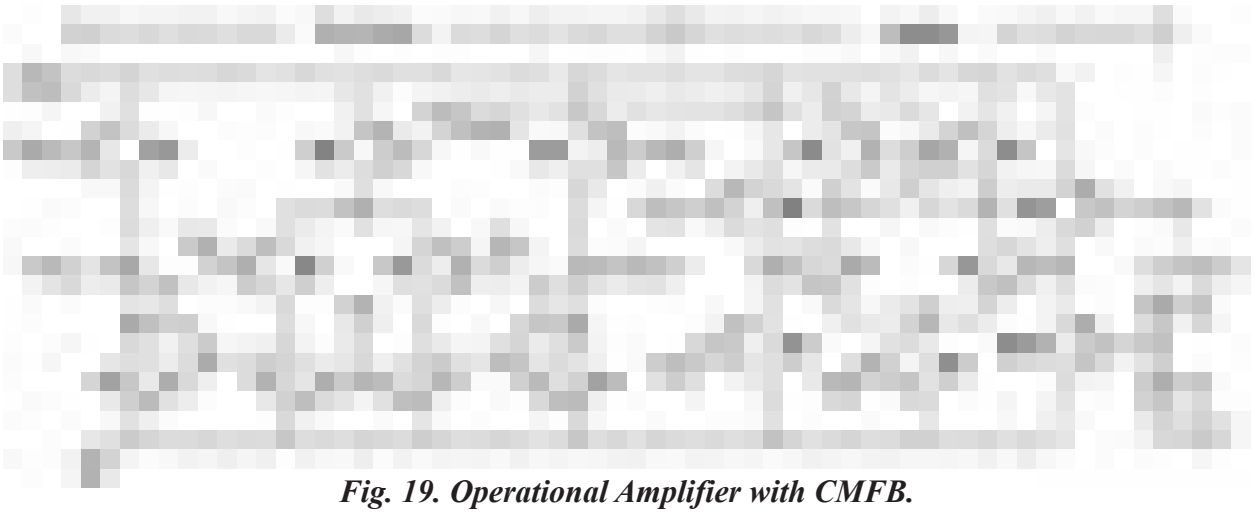


Fig. 19. Operational Amplifier with CMFB.

Bluetooth reception. Each baseband channel has a bandwidth of 5.5MHz. With 2X oversampling, a 44MS/s ADC rate is needed. Each ADC is implemented by interleaving two, 22MS/s sub-ADCs, yielding a subsample period of 45ns. Some key info about this design is shown in Table 4. The opamp design uses MOSFETs in the first stage for high input impedance needed in switched-cap circuits. The use of an HBT in the second stage (Q2) gives high g_m and low input capacitance. This pushes the second pole higher while keeping current low. Also, input stage g_m is multiplied by β to develop voltage gain across the output resistance of M3 (r_{o3}). The high V_A of HBTs should keep the output conductance of Q2 insignificant and the high β of the HBT contributes directly to voltage gain. The approximate no-load gain is:

$$A_{DM} = g_{m1}\beta r_{o3}$$

...where g_{m1} is of M1.

Table 4. Data about the opamp and ADC system.

Opamp open loop gain (A_V)	77dB* (*simulated)
Opamp gain-bandwidth (GBW)	165MHz*
Opamp phase margin	67° *
Supply voltage	2.5V
Differential signal level	2Vpp
Process	0.25um
ADC max sample rate	44MS/s (23ns/S)
ADC resolution	~10-bits (8-11)
ADC power	20mW at 44MS/s
ADC area	2.1mm^2

Differential Mode Dynamics

Positive slew rate is limited by M3 drain current charging the relatively high load capacitance of the following MDAC. Negative slew rate can take advantage of the high sinking capability of Q2 to drive the load. Thus, it will be limited by the drain current of M2 charging C_C. M3 bias current will be set by the slew rate requirement. This fixes the bias current of Q2. Since HBTs need high current density to maximize f_T, Q2 would be sized accordingly. We can write the differential gain-bandwidth as:

$$GBW_{DM} = \frac{2\pi g_{m1}}{C_C}$$

Common Mode Dynamics

The authors state that the bandwidth of the CMFB system should be comparable to the differential mode to insure stability and performance. A bipolar diffamp provides the g_m needed for this while MOS source followers relieve loading effects. Let's trace the CMFB path to develop the common mode GBW expression: Starting at the gate of M5B, assume that the load impedance of the base of Q3 is much higher than $1/g_{m5}$, making the gain of M5B, unity. The collector current of Q3 is generated by $g_{mQ3}/2$, since the diffamp is driven single-ended. Assuming that M7 and M2 are equal width, that current is mirrored into the M1 diffamp where it divides by two and charges C_C, equally on both sides. That voltage arrives at the gate of M5B unattenuated by R_L, since there is no significant load. Thus, we have the common mode bandwidth as:

$$GBW_{CM} = \frac{\pi g_{mQ3}}{2C_C} = \frac{\pi I_{CQ3}}{2C_C V_T}$$

I_{CQ3} is chosen to make this comparable to GBW_{DM}.

Common Mode Summing

We are interested in the resistive divider used to sum the differential outputs. One would expect that, unbuffered, it would reduce differential mode gain unacceptably or require very high (physically large) resistor values. Baker [43, p.891] uses buffered outputs to drive 20K summing resistors. The authors of our opamp do state that it needs >75dB of gain, operating in the S&H of the system. Although they use switched-capacitor CMFB in the MDACs, it must be continuous-time in the S&H. Without the device and current details of the opamp circuit design, we can only guess at the solution hidden behind the apparently simplified summing resistor network. Suppose C_C=1pF. From the 165MHz GBW, we calculate g_{m1}=26μA/V. Using the A_{DM} expression and taking β=200, we plug-in the 75dB needed gain and solve for the total one-sided output load resistance, obtaining r_{out}=1.1MΩ.

Resistors that high would be hard to implement in a typical 0.25μm BiCMOS process. The authors didn't elaborate on how they did this. However, a brief search of the literature revealed that researchers at Texas A&M have published a number of papers related to CMFB, including [70], which shows a circuit which could be used in place of the summing resistors (with modifications). Both Baker [43, pp. 881-895] and Gray [37, pp.828-832] discuss various approaches to the CMFB "detector," as the summing problem has been called.

For any high-resistance summing solution, it is necessary to bypass the resistors with capacitors to maintain the high-frequency response. The capacitors should be sufficiently sized so that the input

capacitance of M5B doesn't unduly degrade the loop bandwidth. Some loss could be allowed, though, to keep the capacitors comparable to the M5B input capacitance.

Bandgap Reference

While the bandgap reference (BGR) can be readily implemented in pure CMOS, we noticed that the BiCMOS design shown in Figure 20 [57] offers exceptionally low noise. Performance results for the circuit are in Table 5. While this 1994 design was targeted at a 5V supply, we should be able to preserve the low noise in lower voltage designs.

Fig. 20. Low noise bandgap reference.

Table 5. Low noise bandgap.

Power	5V, 70uA
Reference Voltage	1.264V
Tempco	9ppm/°C
PSRR	73dB
Noise	12nV/Hz$^{1/2}$
Load Regulation	6-ohms

BGRs are based-on summing V_{BE} with a voltage proportional to V_T. Since V_T rises with temperature (PTAT*) and V_{BE} falls (CTAT*), they can be weighted for a zero tempco at a nominal temperature. The resulting voltage varies about 0.4% over 100°C (40ppm/°C) with a curvature as seen in Figure 21 [37, edited].

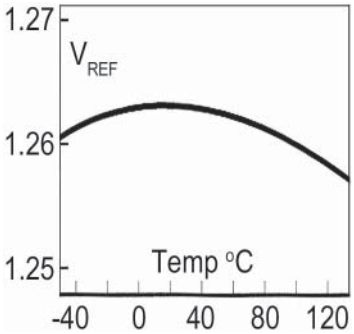


Fig. 21. Basic bandgap temperature drift.

In Figure 20, the loop of Q2, MP2, MP1 and Q1 generate a PTAT current. MP1-4 form a current mirror transferring Q2 collector current (I_{Q2}) to Q1. The Q2 emitter is made 4X the size of Q1. (Ignore Rcomp since it just evens-up the base resistance of Q1 and Q2). From [34] we can express the difference in V_{BE} between Q1 and Q2 as: $\Delta V_{BE} = V_T$ **ln**(4) = 1.4V_T. This voltage appears across R1, making I_{Q2} vary, PTAT. I_{Q2} (same as I_{Q1}) is cloned by Q3, converted to a voltage by R2 and summed with the CTAT V_{BE} of Q4 to form the stable reference voltage.

So far, this is just the story of the original bandgap reference with 40ppm/°C stability. What [57] brings to the table is that the base current of Q4 can be used to compensate the predictable curvature of the basic BGR. To do this, MP5 is sized to set Q4's base current appropriately to effect the compensation. The authors give a complex but tractable expression for $V_{BE}(T)$. The reference circuit produces this voltage function of temperature:

$$V_{REF}(T) = V_{BE}(T) + K_1 T + K_2 T \exp\left(\frac{\Delta E_G}{kT}\right) \qquad \beta_\infty, \Delta E_G \text{ - measurable process parameters}$$

The design procedure is to use numerical optimization to find K_1 and K_2 for lowest drift. We set:

$$R_2 / R_1 = \frac{qK_1}{k \ln(A_2 / A_1)}$$ where A2/A1 is the emitter ratio of Q2/Q1.

---

* PTAT stands for, "Proportional to Absolute Temperature."
CTAT stands for, "Complementary to Absolute Temperature."

Then the widths of MP5 and MP2 are ratioed according to:

$$W_{MP5}/W_{MP2} = 1 + \frac{K_2\beta_\infty}{\left(\frac{kR_2}{qR_1}\right)\ln(A_2/A_1)}$$

The results of using the technique are shown in Figure 22. While other compensation methods were previously reported, the authors point out that their approach is simpler.

Fig. 22. Results of using the curvature compensation scheme of [57].

Of course, this scheme is based-on BJT characteristics so it is not applicable to pure CMOS designs. Other advantages which the BJT brings here are the low noise and low offset voltage. In CMOS BGRs, offset voltage is usually the largest source of drift [37, p.325].

Finally, as was examined in the section, "The Strange Case of the HBT Emitter," the oxide in the emitter between the poly and mono silicon has a significant effect on β and its temperature coefficient. Hence, one should expect that this curvature compensation technique can be affected by the properties of that oxide in a particular process.

Pipeline-Flash Analog-to-Digital Converter (ADC)

Earlier, we discussed the opamp used in the S&H of the ADC in [1]. Here we will examine the pipeline ADC in this remarkable chip. The design is impressive due to its extraordinary breadth of capabilities and the fact that it is presented by just one student and his professor. (Maybe there were others, since the support of five others was acknowledged.)

We call its breadth "extraordinary," because it embodies the following range of high-performance subsystems and features (some embedded within others listed):

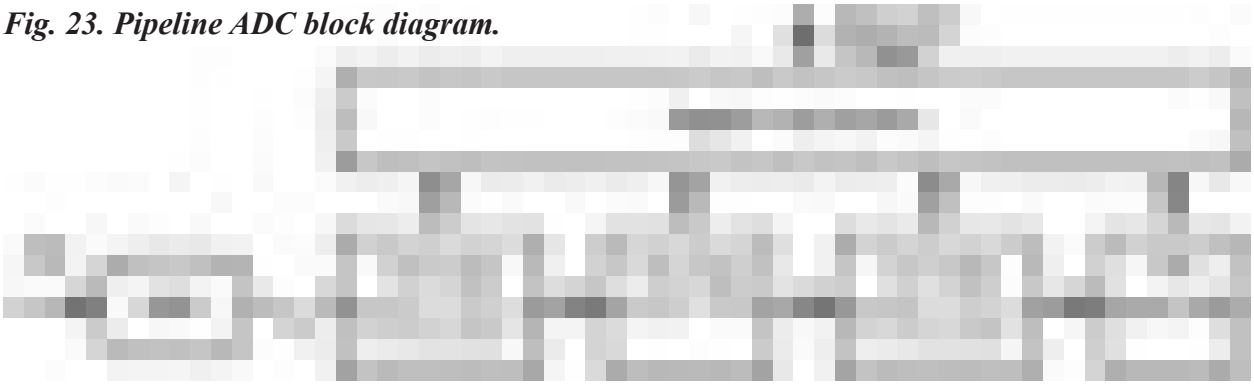
- Differential opamp and S&H with CMFB (discussed earlier).
- 10-bit 44MS/s dual/quad-interleaved ADC.
- Digital nonlinearity correction.
- Capacitor-based 3-4-bit MDACs corrected to 10-bits.
- Dual-mode Bluetooth and 802.11b RF frontend.
- High-speed, low-power, low-offset comparator.
- ~40dB operating range (estimated from 60dB SNR).
- Dual-frequency analog baseband filters.
- 20mW power with 2.5V supply.
- 11mm^2 area in 0.25μm BiCMOS.

While there is much to appreciate in this design, we will focus on two aspects:

- BiCMOS architecture considerations in the ADC.
- Circuit design of the pipeline-flash ADC stage and comparator.

For our purposes, we will burrow into the chip to a single ADC subsystem as shown in Figure 23 [1, edited]. This ignores the RF frontend, the baseband subsystems, the S&H, the secondary ADCs and the interleaving apparatus. It will be visualized in the Bluetooth (BT) mode, which uses a single pipeline in the ADC.

Fig. 23. Pipeline ADC block diagram.



In the pipeline, each of the four stages performs the function seen in Figure 24 [1, edited]. The analog input signal is converted to three or four bits by the small, flash ADC. That value is converted back to analog by the DAC and is subtracted off of the input. The remainder is amplified for the next stage. Since following stages convert the remainders, the ADC values from all the stages, when concatenated, yield a proper conversion having the composite wordlength.

In a pipeline ADC, the last stage only needs to have the accuracy of its individual wordlength. However, the first ADC must be able to define its LSB thresholds to the accuracy of the composite wordlength.

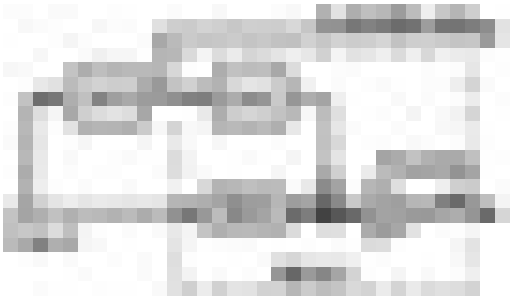


Fig. 24. One stage of the pipeline.

BiCMOS Architecture Considerations of the ADC

Generally, CMOS data converters tend to use capacitor-based weighting circuits while BJT converters use resistor weighting. Capacitors work well with the high impedance gates and capitalize on the excellent switching capability of CMOS. On the other hand, BJT current cells using emitter degeneration resistors show excellent current matching, limited mainly by the resistors [36, p.338]. Thus, it's interesting to see the choices made between the two approaches in the subsystems of this chip. The DAC in the pipeline stage is implemented using switched-capacitor weighting techniques. However, we do not know of a capacitor-based equivalent to the flash ADC.

Fig. 25. Basic Flash ADC.

A flash ADC is basically a resistor string with comparators at each node as illustrated in Figure 25 [36]. Bipolar reference buffers excel at providing the low-offset buffering needed to drive the resistor string. They can also provide high g_m at low area and power needed in the comparators.

Differential Flash ADCs

In Figure 26, we have a comparator from the flash ADC stage. Since this is a fully differential design, there are two input buses (V_{IN+},V_{IN-}) and two resistor string buses. To visualize the differential resistor strings, imagine two strings like the one in Figure 25, except one is tied to $-V_{REF}$ and the other to $+V_{REF}$. Let's call the voltage at a particular node on the string, V_R. Now, each comparator is supplied with $-V_R$ and $+V_R$ whereas the single-ended version only gets $+V_R$. You might think that having two reference voltages instead of one is redundant but it isn't really. While $+V_R$ and $-V_R$ are the same magnitude, they are, of course, opposite in sign. Maintaining this symmetry around ground maximizes the cancellation of common-mode noise and cancels even-order nonlinearity.

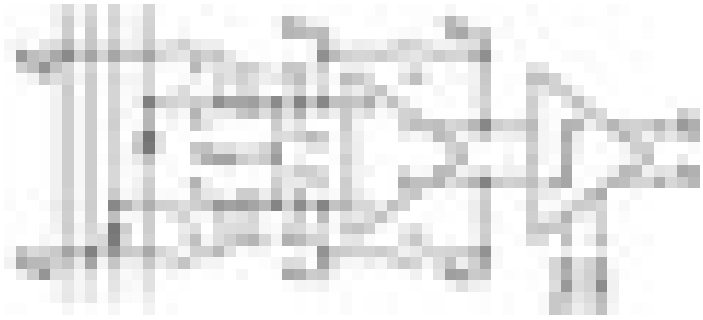


Fig. 26. Comparator stage of the flash ADC.

The conceptual differential resistor string uses dual polarity V_{REF}'s but these days, analog circuits are often powered from a single supply. A common-mode reference voltage (V_{CM}) is established which serves the purpose that ground does for dual-supply systems. V_{CM} is between the positive rail and ground. It is sometimes set at $V_{DD}/2$ but in an ADC, a better choice may be $V_{REF}/2$. We can use the dual resistor strings to define $V_{CM}=V_{REF}/2$.

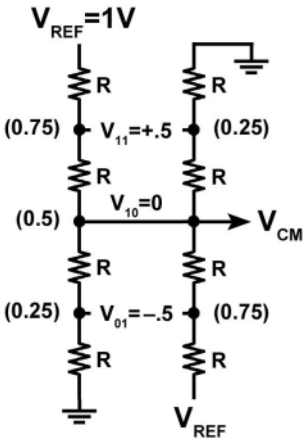


Fig. 27. Differential resistor strings for two-bit flash ADC.

An example pair of resistor strings for a two-bit, single-supply ADC is shown in Figure 27. The values are for an assumed reference voltage of 1.0V. In the middle column are the differential voltages presented to the comparators. There will be three comparators, dividing the ±1V differential signal range into four voltage ranges. The ADC would produce the codes in the table at right for those ranges.

V_{IN} (diff)	Binary Out
-1.0V to -0.5V	0 0
-0.5 to 0.0	0 1
0.0 to 0.5	1 0
0.5 to 1.0	1 1

The Importance of Nothing (0V)

Notice that the comparator in the middle of the strings has both reference inputs tied to V_{CM} and the differential reference value is zero. That is what we wanted, since we designed the ADC to have a transition at 0V signal level. There is no code which represents 0V. We do have codes for +0.25V and –0.25V, which are the centers of the 10_2 and 01_2 ranges.

There are reasons though, why we might prefer *not* to have a transition at zero. First, it means that very low level noise can be promoted to the level of 1-LSB. This can increase sensitivity to digital noise on the chip and thermal noise in the analog circuits. Second, low level sine waves, (e.g. 60Hz hum) can be converted to square waves of 1-LSB. The increase in distortion of these signals may make them objectionable. Finally, it is nice to have a code for zero so that when we apply 0V to the ADC, we get a known output and can verify that the input offset is low.

While these issues may not be major, they are enough to push most designs to balance the two LSB thresholds around zero. The downsides of this are (1) that we have to insert four extra R's of

value R/2 in the strings and (2) there will be one extra code on one polarity of V_{IN}, than the other. For details on the four R's, please refer to [2]. The extra code on one side doesn't sound like much but it could be a consideration for short word length ADCs used in pipelines like the one in Figure 23. As a practical matter, at 3-bits, it's only 12% difference, so it probably doesn't have much effect. For 2-bit ADCs with a zero code though, there is only one negative code. One code is for zero and the other two are positive.

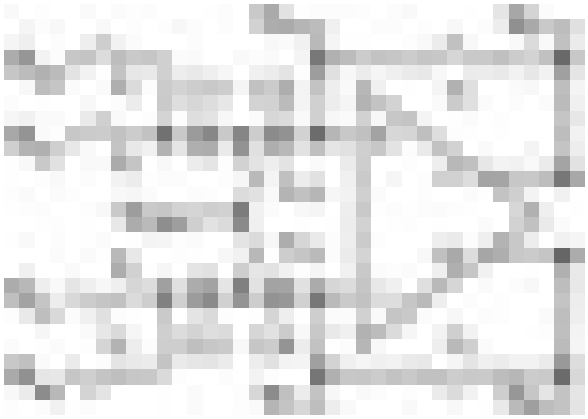


Fig. 28. Differential switching comparator.

Dynamic Comparator Operation

We have taken the preceding diversion to better understand what the comparator in the flash ADC has to do. Getting back to the one in Figure 26, we now know that it must have four inputs: a differential pair for the input signal and another pair for the reference level. The comparator must decide which of the two differential values is greater. In this design, it does this on a cyclic basis using four clock phases. Let's look at the comparator switching circuits, using the enlargement in Figure 28 [1, edited] and the timing in Figure 29 [1, edited]. The four clock phases, ϕ_1, ϕ_2, ϕ_{2e}, ϕ_3 are all the same frequency and operate the switches as labeled. Since ϕ_2 is almost the same as ϕ_{2e}, let's refer to them both as ϕ_2 for now and do a tweak later. We will think of each clock signal in terms of when it is asserted; that is, the time when its switches are closed. Notice that the switches operate in pairs, so we only have four switch pairs to consider, and two of them operate on ϕ_2.

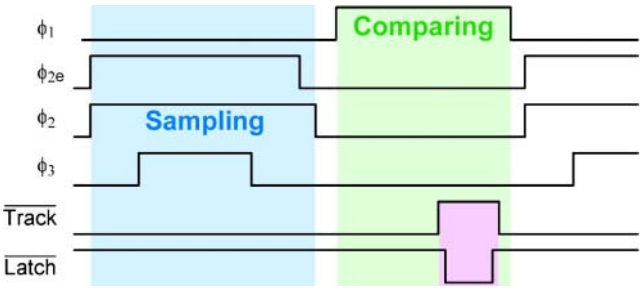


Fig. 29. Comparator clocking. H=switch closed.

There are two main periods of operation during the clock cycle: Sampling and Comparing. Sampling acquires the reference value and the preamp offset. In the Sampling period, all switches except ϕ_1 close. With ϕ_2 closed, C1 and C3 sample V_R. With ϕ_3 closed, C2 and C4 sample the offset voltage of the comparator preamp. (We will see that it is effectively the comparator offset.) Comparing renders the result. While Comparing, ϕ_1 is closed and all the others are open. The comparator preamp then sees $V_{IN} + V_R - V_{OFF}$. The amplified result is presented to the comparator and the result is latched. The Sampling and Comparing periods do not overlap.

Another thing about the Sampling period: ϕ_2 closes before ϕ_3 and stays closed later than ϕ_3. That isolates the inputs from the stage by connecting the caps to V_{CM}. This prevents disturbance (known as kickback) of the preceding circuits by the comparator.

Bottom Plate Sampling

To complete the Sampling period discussion, let's look at ϕ_2 versus ϕ_{2e}. They are almost the same but ϕ_{2e} ("ϕ_2 early") reopens before ϕ_2 does. (We will use the C1/C2 side for discussion.) As explained by Baker [43, p.841], the reason for this order is that the charge injected by ϕ_{2e} is independent of signal amplitude, since the S/D are at V_{CM}. Whereas ϕ_2 charge injection varies with

the instantaneous voltage present on its S/D when it switches. So ϕ_{2e} switches first, injecting a fixed offset. With ϕ_{2e} open, the right side of the C1~C2 string is open (except for the opamp input). The source impedance of V_R is lower and provides a path for the charge injection. Little charge is injected in C1 and C2. Baker also mentions that the bottom plate of C1 should be on the left, allowing the source impedance of V_R to shunt much of the substrate noise coupled to the bottom plate. He calls the phasing and orientation technique, "bottom-plate sampling."

Fig. 30. Preamp-comparator-latch circuit.

Preamp and Input Stage

Now let's see what is happening with the internal comparator operation. For this, we drill into the preamp/comparator/latch circuit in Figure 30. The preamp is needed to buffer the small input capacitors from the loading of the comparator's bipolar input stage. With its high g_m at low area and power, the bipolar stage causes fast switching for small input voltage differences. As we saw earlier, the HBTs have high f_T and V_A. The f_T supports the switching speed and the V_A increases CMRR. Intuition might lead you to think that the higher CMRR of the bipolar stage would be spoiled by the CMRR of the preamp (as offset is). But as Baker noted [43, p.823], the CMRR of cascaded diffamps is the *product* of their CMRRs. As a result, cascaded diffamps have a major advantage in that respect. For example, mediocre CMRRs of say, 35dB and 45dB, will become a respectable 80dB, when cascaded.

Now, since the V_{OFF} of the bipolar stage is small compared to that of the preamp, we need not be concerned with getting lots of voltage gain in the preamp. The offset zeroing which is performed during the Sampling period will be used to deal with the high offset of the MOS diffamp. A modest amount of preamp gain should be able to reduce the input-referred V_{OFF} of the bipolar stage to less than that of the residual charge injection. Accordingly, the preamp drain load "resistors" can be provided by just weak pMOSFETs. We are not much concerned with their nonlinearity because once the comparator gets past the decision point, the exact magnitude of the response isn't critical. Being just a single stage, the stability of the preamp with unity feedback is assured without any compensation.

Notice how simple the preamp's ϕ_3 feedback is. Operating the input FETs at V_{DG}=0V assures operation in the saturation (high r_o) region since (for long channels) $V_{DSsat} = V_{GS} - V_{TH}$ and here

$V_{DS}=V_{GS}$. (For short channels, V_{DSsat} is even lower.) So the unloaded gain of the preamp becomes:

$$A_{Vpre} = \frac{g_{mN}}{g_{mP}} = \sqrt{\frac{KP_N W_N L_P}{KP_P W_P L_N}} \approx 1.7\sqrt{\frac{W_N L_P}{W_P L_N}} \qquad \text{(Pegging the KP ratio at 3.)}$$

Regenerative Comparator and S-R Latch

The regenerative comparator is where the track/latch action happens. "Latch" is pretty self-evident but the term "track" is a bit more obscure for a comparator. We can see in Figure 30 that when Track_bar is asserted low, the outputs of the bipolar diffamp are connected together by M3. This eliminates the regeneration between the diffamp outputs which is provided by the pMOSFET loads. During Track mode, then, the bipolar diffamp currents are controlled by its differential input voltage and its differential output voltage is zeroed. The collector loads are approximately $1/g_{mP}$ on each side.

Thus, its output *voltage* as Track_bar becomes de-asserted, is unbiased by previous signals. At that point, though, the bipolar diffamp *currents* differ by the effect of the input voltage differential. One collector begins moving lower than the other. (Let's say it's the M1 side.) Regeneration between M1 and M2 quickly reinforces the decision, switching to a stable state. The M2-side collector can go nearly to the positive rail (2.5V) as M2 turns on. The M1 side turns off, allowing the collector to pull down to its limit. We can write the worst-case limit as:

$$V_{CL} = V_{DD} - V_{THP} - V_{BE} + V_{CEsat}$$

Considering I_C will be minimal, we estimate the collector low level as:

$$V_{CL} = 2.5 - 0.5 - 0.7 + 0.05 = 1.35V \text{ *}$$

Up to now, Latch_bar has been high, keeping M4 and M5 off. It has also kept M6 and M9 on, forcing the outputs of the regenerative comparator to zero. When Latch_bar is asserted low, M6 and M9 turn off, releasing the outputs. M4 and M5 turn on, coupling the collector voltages to cross-coupled pair, M7 and M8. In effect, we have two inverters, M1/M7 and M2/M8, driving each other. This second level of regeneration pulls the lower side down to ground and leaves the high side near V_{DD}. Full CMOS levels are presented to the next stage.

The S-R latch is two, standard, cross-coupled CMOS NOR-gates. When both of its inputs are low (before Latch_bar is asserted), it maintains its current state. After Latch_bar is asserted, one of the outputs of the regenerative comparator goes high, transferring the decision to the latch. Notice that Latch_bar is de-asserted before Track_bar is reasserted, protecting the latched data. Track_bar then becomes asserted and the bipolar stage goes back to Track mode.

---

* V_{THP} is from [3].

What Happens When Supplies Drop Below 1V?

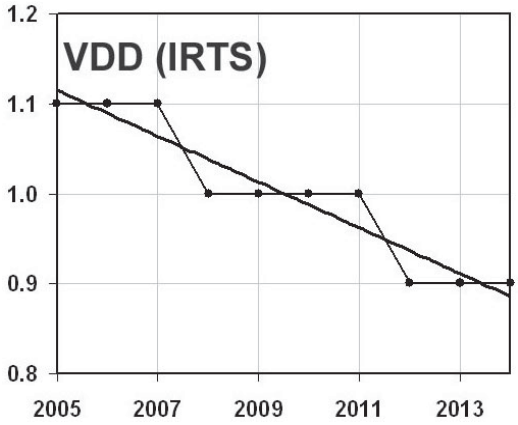


Fig. 31. CMOS power supply roadmap.

From the IRTS roadmap [7] we have the projections of integrated circuit power supply trend shown in Figure 31. That indicates that 1.0V V_{DD} is expected in 2008. Indeed, from a plot of first-shipment dates in [17], the 1.0V threshold was crossed in 2003. As we saw in our section on HBT transistor characteristics, the increased base doping raises V_{BE}. Its typical operating values run as high as 0.92V. This makes it very difficult to operate HBT circuits with V_{DD} as low as 1V. Widlar [32] has shown that typical BJTs can be useful as low as 1.1V, albeit with low signal swing. Kimura [33] also has shown BJT reference circuits operating down to that level.

However, from the trend, circuits soon will be required to operate at less than 1.1V. What does this mean for the future of the HBT? Here are some *possible* answers to that question:

(1) The HBT will disappear. (?)

Well, this would be a big surprise to IBM, which is now on its fourth generation of HBT process development. Perhaps the most telling rebuttal to this answer is in Figure 32 [28]. If we "follow the money," we can see that the industry has good reason to keep the HBT around.

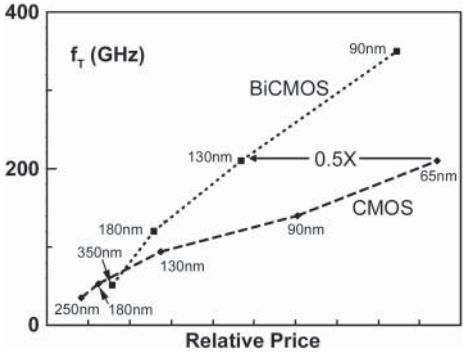


Fig. 32. CMOS and BiCMOS f_T performance versus price.

To argue the other side, Critchlow [17] comments:

> "The 2.5V CMOS Technology was the death knell for high performance silicon bipolar technologies in high-end computers. BiCMOS had gathered some momentum, but when designers came to realize that very effective off-chip drivers could be made using MOSFET circuits, BiCMOS soon faded."

Could this be pre-HBT thinking? Is it the view from the high performance digital side? We suspect it is the latter. From the perspective of consumer communications products (e.g. cellphones) the HBT has advantages and cost is critical. Thus, this comment might not be opinion that HBT will disappear altogether but that CMOS will take over its role in some digital areas.

(2) Very low voltage may not be as important in portable applications.

We noticed a comment in the literature about the difficulty in following the scaling progression with portable applications. The problem is that single-cell battery voltages are generally around 1.2V or greater, which tends to oppose scaling below that bar. This has a similar bearing on the HBT power issue. Even if a switching converter is used to power the CMOS at lower voltages, there would still be a source of higher voltage available without additional cost. Thus, the use of a second supply for HBTs would not have a penalty (other than an extra pin). Since portable devices are a significant and cost-sensitive part of the market, one would expect HBTs to be used so long as they have a cost advantage.

(3) Multiple supplies are an accepted way to support BiCMOS chips.

Designers at the PCB level are quite comfortable with the need to have multiple power supplies. With today's plethora of chips of varying levels of process advancement, it is common for small switchers to be included in a design. For example, at my last employer, one modest-size PCB had a bank of six switching regulators. The cost was less than $1 each. A recent consumer video decoder chip from Sigma Designs, SMP8630, requires three supplies: 3.3V, 2.5V and 1.2V.

Conclusions

In this paper, we have seen that BiCMOS technology has advantages in terms of noise, offset voltage, bandwidth, impedance matching, transconductance, voltage gain, ESD survival and cost. On the other hand, we have also seen that CMOS designers have found ways to mitigate many of these issues, so that pure CMOS does make a formidable competitor for BiCMOS.

Deciding which is better requires an analysis of the particular application. Generally, all-digital chips would tend to be better served by CMOS whereas RF mixed-signal designs would generally do better in BiCMOS.

One thing is clear: The heterojunction bipolar transistor (HBT) is a *de facto* standard in BiCMOS processes. The enhanced β, f_T, and Early Voltage account for much of the BiCMOS advantage. When the HBT appeared on the scene around 1995, BiCMOS was apparently waning, at least for digital applications. The HBT rescued the BJT and has not lost ground against the MOSFET, in terms of f_T, in the three process generations which followed.

In our circuit design examples, we have seen that the BJT can play an important role in reducing power, decreasing noise, increasing reference stability and optimizing flash analog-digital converters. While the actual number of BJTs employed may not be large, their contribution of high g_m at low power and size, along with low noise and offset, can greatly benefit otherwise CMOS designs.

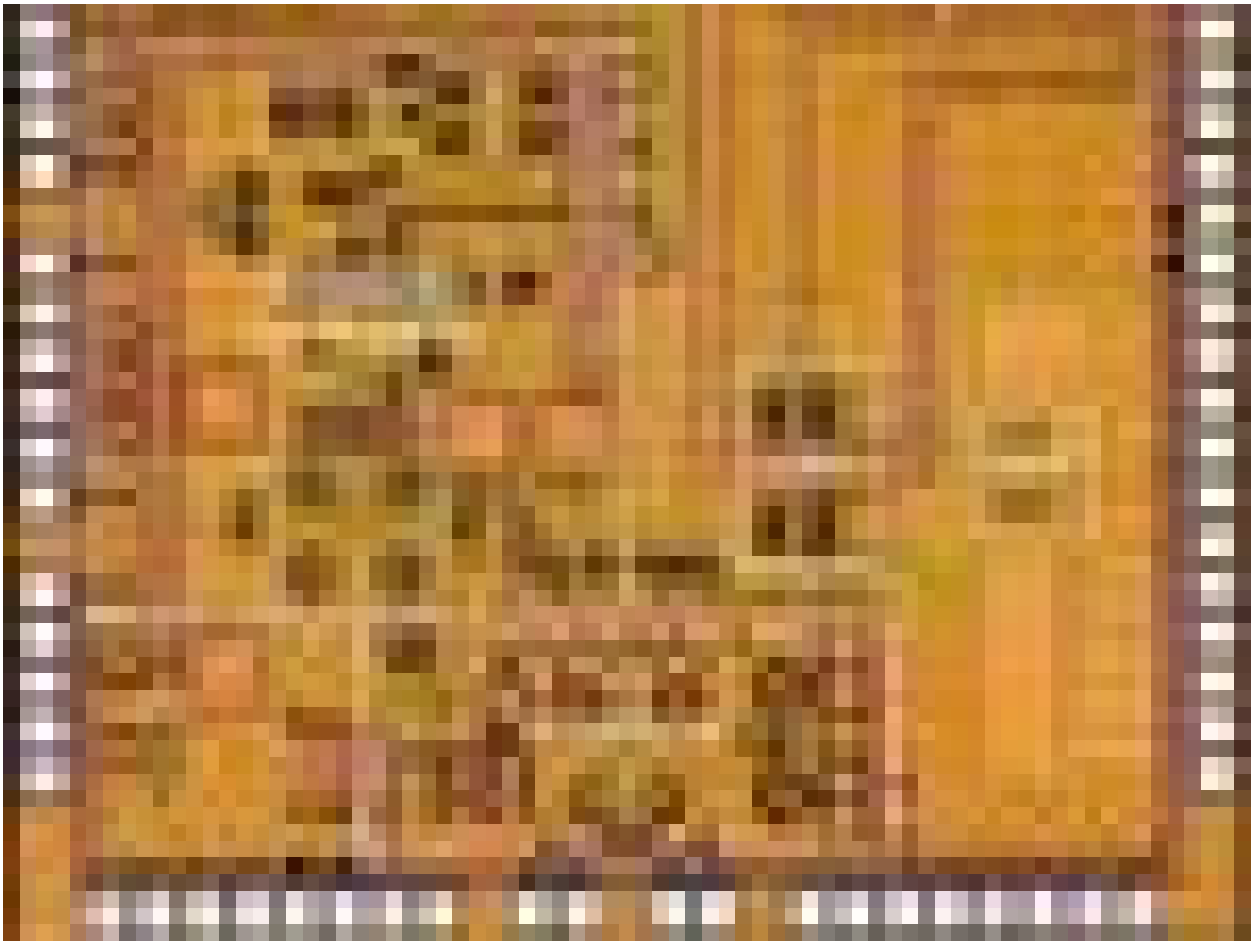
The Unanswered Questions

Finally, looking back at the questions which were posed in the Introduction, we are satisfied that they have been well covered, except for one (two-parter): "Why didn't Baker address BiCMOS? Why did Gray's 4th edition cover bipolar and CMOS separately, except for several pages?" Of course, we would not presume to speak for these great authors, so we will only try to identify what we think are plausible reasons.

Taking Gray's book first: Now in its fourth edition, he has extended this great body of work. It originated in 1977 with barely a mention of MOSFETs, so it would be natural to treat them separately. In the BiCMOS designs we have seen, either BJT circuits are mostly separate (as in RF frontends) or a small number of BJT devices are employed in strategic places in otherwise CMOS circuits. It is not so common to ingeniously combine scores of BJTs and MOSFETs, mixed in exotic ways. Thus, there is less reason to treat them in an integrated fashion.

As for Baker not addressing BiCMOS: It is a great challenge for CMOS analog circuit designers to overcome limitations without having to fall back on the crutch of a process complication. Designers are delighted by finding elegant solutions which pit their skill against the vagaries of the process. With scaling now deep in the submicron range, MOSFETs offer seemingly limitless quantity to offset their limitations in certain qualities. This almost demands that complex and clever circuit designs be created—an irresistible temptation for a great circuit designer.

While the outlook for BiCMOS seems bright, the fact that in a way, it is pitted against the skill and ingenuity of CMOS circuit designers gives one a little disquiet about its future.



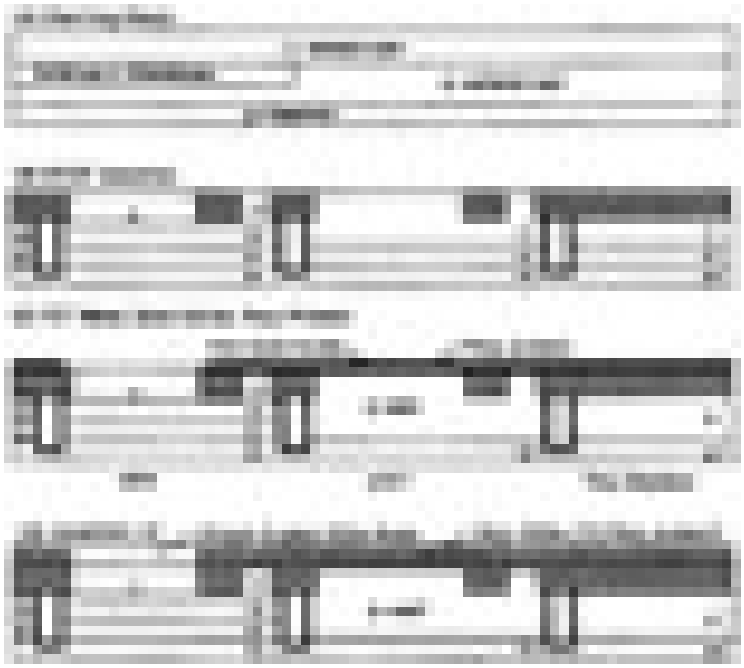
Ultra-wideband receiver in SiGe BiCMOS (Texas A&M). *(decorative)*

Appendix A – Baker's 50nm NMOS SPICE Model [47].

```
*** Short channel models from CMOS Circuit Design, Layout, and Simulation, 2e
* 50nm BSIM4 models Vdd=1V

.model  N_50n  nmos  level = 14

+binunit = 1            paramchk= 1            mobmod  = 0
+capmod  = 2            igcmod  = 1            igbmod  = 1            geomod  = 1
+diomod  = 1            rdsmod  = 0            rbodymod= 1            rgatemod= 1
+permod  = 1            acnqsmod= 0            trnqsmod= 0

+tnom    = 27           toxe    = 1.4e-009     toxp    = 7e-010       toxm    = 1.4e-009
+epsrox  = 3.9          wint    = 5e-009       lint    = 1.2e-008
+ll      = 0            wl      = 0            lln     = 1            wln     = 1
+lw      = 0            ww      = 0            lwn     = 1            wwn     = 1
+lwl     = 0            wwl     = 0            xpart   = 0            toxref  = 1.4e-009

+vth0    = 0.22         k1      = 0.35         k2      = 0.05         k3      = 0
+k3b     = 0            w0      = 2.5e-006     dvt0    = 2.8          dvt1    = 0.52
+dvt2    = -0.032       dvt0w   = 0            dvt1w   = 0            dvt2w   = 0
+dsub    = 2            minv    = 0.05         voff1   = 0            dvtp0   = 1e-007
+dvtp1   = 0.05         lpe0    = 5.75e-008    lpeb    = 2.3e-010     xj      = 2e-008
+ngate   = 5e+020       ndep    = 2.8e+018     nsd     = 1e+020       phin    = 0
+cdsc    = 0.0002       cdscb   = 0            cdscd   = 0            cit     = 0
+voff    = -0.15        nfactor = 1.2          eta0    = 0.15         etab    = 0
+vfb     = -0.55        u0      = 0.032        ua      = 1.6e-010     ub      = 1.1e-017
+uc      = -3e-011      vsat    = 1.1e+005     a0      = 2            ags     = 1e-020
+a1      = 0            a2      = 1            b0      = -1e-020      b1      = 0
+keta    = 0.04         dwg     = 0            dwb     = 0            pclm    = 0.18
+pdiblc1 = 0.028        pdiblc2 = 0.022        pdiblcb = -0.005       drout   = 0.45
+pvag    = 1e-020       delta   = 0.01         pscbe1  = 8.14e+008    pscbe2  = 1e-007
+fprout  = 0.2          pdits   = 0.2          pditsd  = 0.23         pditsl  = 2.3e+006
+rsh     = 3            rdsw    = 150          rsw     = 150          rdw     = 150
+rdswmin = 0            rdwmin  = 0            rswmin  = 0            prwg    = 0
+prwb    = 6.8e-011     wr      = 1            alpha0  = 0.074        alpha1  = 0.005
+beta0   = 30           agidl   = 0.0002       bgidl   = 2.1e+009     cgidl   = 0.0002
+egidl   = 0.8

+aigbacc = 0.012        bigbacc = 0.0028       cigbacc = 0.002
+nigbacc = 1            aigbinv = 0.014        bigbinv = 0.004        cigbinv = 0.004
+eigbinv = 1.1          nigbinv = 3            aigc    = 0.017        bigc    = 0.0028
+cigc    = 0.002        aigsd   = 0.017        bigsd   = 0.0028       cigsd   = 0.002
+nigc    = 1            poxedge = 1            pigcd   = 1            ntox    = 1

+xrcrg1  = 12           xrcrg2  = 5
+cgso    = 6.238e-010   cgdo    = 6.238e-010   cgbo    = 2.56e-011    cgdl    = 2.495e-10
+cgsl    = 2.495e-10    ckappas = 0.02         ckappad = 0.02         acde    = 1
+moin    = 15           noff    = 0.9          voffcv  = 0.02

+ktl     = -0.21        ktl1    = 0.0          kt2     = -0.042       ute     = -1.5
+ual     = 1e-009       ubl     = -3.5e-019    ucl     = 0            prt     = 0
+at      = 53000

+fnoimod = 1            tnoimod = 0

+jss     = 0.0001       jsws    = 1e-011       jswgs   = 1e-010       njs     = 1
+ijthsfwd= 0.01         ijthsrev= 0.001        bvs     = 10           xjbvs   = 1
+jsd     = 0.0001       jswd    = 1e-011       jswgd   = 1e-010       njd     = 1
+ijthdfwd= 0.01         ijthdrev= 0.001        bvd     = 10           xjbvd   = 1
+pbs     = 1            cjs     = 0.0005       mjs     = 0.5          pbsws   = 1
+cjsws   = 5e-010       mjsws   = 0.33         pbswgs  = 1            cjswgs  = 3e-010
+mjswgs  = 0.33         pbd     = 1            cjd     = 0.0005       mjd     = 0.5
+pbswd   = 1            cjswd   = 5e-010       mjswd   = 0.33         pbswgd  = 1
+cjswgd  = 5e-010       mjswgd  = 0.33         tpb     = 0.005        tcj     = 0.001
+tpbsw   = 0.005        tcjsw   = 0.001        tpbswg  = 0.005        tcjswg  = 0.001
+xtis    = 3            xtid    = 3

+dmcg    = 0e-006       dmci    = 0e-006       dmdg    = 0e-006       dmcgt   = 0e-007
+dwj     = 0.0e-008     xgw     = 0e-007       xgl     = 0e-008

+rshg    = 0.4          gbmin   = 1e-010       rbpb    = 5            rbpd    = 15
+rbps    = 15           rbdb    = 15           rbsb    = 15           ngcon   = 1
```

Appendix B – A BiCMOS Process

This process flow is taken from [59] and is of 1995 vintage. It is identified as IBM's ETx process. The ETx processes mentioned in the paper dated from 1993. They were reporting HBT performance in terms of f_{max} and the highest value mentioned was 60GHz. The process description is also taken from [59], heavily edited for brevity.



The starting wafers (Fig. A) are p- epi on p+ substrates. A patterned n+ subcollector is formed, followed by a thin n- epitaxial layer grown everywhere. Isolation is provided by Polysilicon-filled deep Trench (PST) and oxide-filled shallow trench isolation is then exercised. The planarized isolation is used to achieve a topography free surface. The structure at this point is shown in Fig. B.

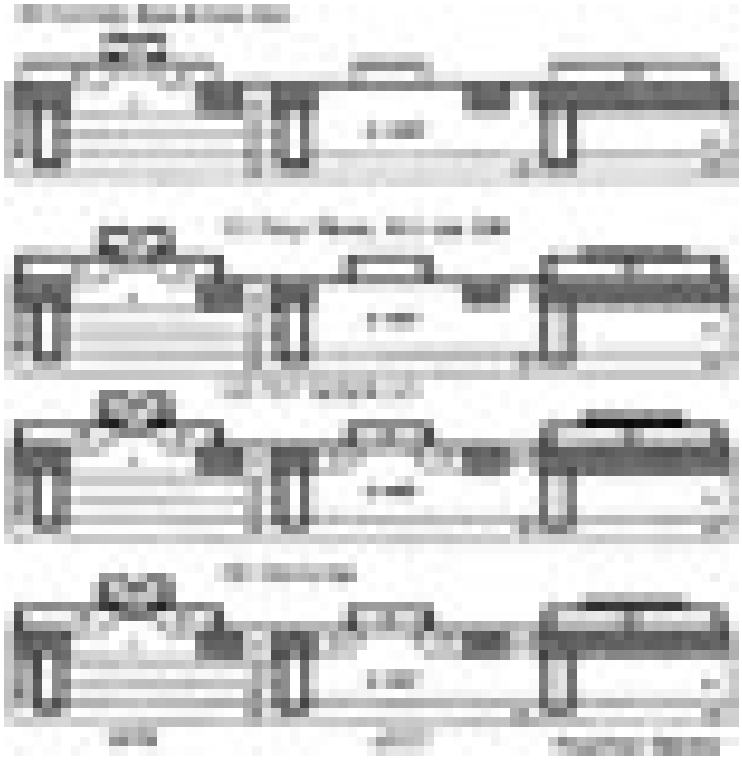
Contact to the n+ subcollector from the top surface is achieved with a deep phosphorus diffusion. This is followed by the n-well and p-well implants. A decoupling capacitor is formed by growing a thin oxide over the n+ reach-through region and patterning an additional deposited nitride layer. A gate oxide of 7 nm is grown and a thin gate-protect polysilicon layer is deposited. The gateprotect polysilicon layer is patterned and removed over the npn active area (see Fig. C).

The silicon or SiGe epitaxial base is grown by the UHV/CVD low temperature epitaxy. Deposition from SiH_4 in UHV/CVD naturally results in polysilicon over field regions and epitaxial silicon over active device regions as shown in Fig. D.

The polysilicon over field regions will serve as a natural extrinsic base electrode. The polysilicon layer serves as the CMOS gate electrode. Following base deposition, a passivation oxide layer is grown. Next a thin layer of nitride is deposited, followed by a thin polysilicon "conversion" layer. The emitter "mandrel" is then completed with deposition of a nitride and a relatively thick oxide layer. The emitter window is defined, stopping on the conversion layer polysilicon as shown in Fig. E.

An oxide formed on the mandrel serves to space a high dose extrinsic base boron implant away from the emitter opening. A resist mask protects the CMOS devices. For the SiGe HBT the implant goes directly into the strained layer. The "topography" required for self-aligned device features is eliminated by wet etch removal of oxide, leaving only a small nitride pad to define the final emitter window. Once the emitter stack has been removed, HIPOX oxidation completely converts the polysilicon over the implanted extrinsic base and field regions to oxide using the nitride emitter pedestal as a local mask. The nitride may now be removed, exposing the unconverted polysilicon plug. Emitter window formation is completed through removal of the polysilicon plug and underlying nitride pad using the "converted" oxide as a hard mask as shown in Fig. F.

The emitter polysilicon is deposited and ion implanted with arsenic, capped with silicon nitride, and patterned with an etch that stops on the dielectrics over the base region (see Figs. F and G).



The extrinsic bases, CMOS gates, poly-resistors, and decoupling capacitors, are all patterned and etched together after the emitter polysilicon. For the CMOS gates, the etch must be highly anisotropic and selective to the gate oxide (see Fig. H).

This requirement also results in a precision resistor and extrinsic base etch. At this point in the process, the emitter is doped with implanted arsenic and the extrinsic base, polysilicon resistors, and decoupling capacitors are all doped with implanted boron, but the FET source, drain, and gate (S/D/G) remain undoped. The emitter remains capped with nitride to prevent dopant loss and prevent silicidation. An oxygen ambient is used for the emitter furnace anneal to reoxidize the gate and extrinsic base sidewalls (see Fig. I). The nitride cap over the emitter protects it from oxidation.

A resist mask is used to pattern the nitride over the polysilicon resistors to block silicide and define the resistor value. A nitride sidewall is formed over the emitter, base, gate, resistor, and decoupling capacitor polysilicon. A dual poly process which simultaneously dopes the S/D/G is used for both FET devices (see Fig. J).

Shallow junctions (≤13ì m) are achieved by preamorphization with heavy ions (Ge for pFET, Sb for nFET), and implanting the dopant species at low energy. Self-aligned titanium disilicide is formed over the extrinsic base poly, poly-resistor contacts, gate poly, and source/drain diffusion areas. Salicidation of the extrinsic base is applied (see Fig. K). Finally, contacts and metallization finish the process.

References

[1] B. A. Valdes-Garcia and E Sánchez-Sinencio. "A 10-bit 44-MS/s 20-mW Configurable Time-Interleaved Pipeline ADC for a Dual-Mode 802.11b/Bluetooth Receiver." *IEEE Journal of Solid-State Circuits,* Vol. 41, pp. 530-539, March 2006.

[2] T.-S. Lee, L.-D. Luo, and C.-S. Lin, "Design Techniques for A CMOS Low-Power Low-Voltage Fully Differential Flash Analog-to-Digital Converter." *47th IEEE International Midwest Symposium on Circuits and Systems,* pp. I-357-360, 2004.

[3] M. Aquilino. "Development of a Deep-Submicron CMOS Process for Fabrication of High Performance 0.25 mm Transistors." Masters Thesis slides, Rochester Institute of Technology, 2006.

[4] Infineon. "Custom RF" web page with "Application - Technology Match" graphic. http://www.infineon.com/upload/Image/WGaal/Grafik_Technology_Features_L90web.jpg, Accessed November 2006.

[5] Infineon. "C11RF Technology Brochure." (C11RF-br-2006.pdf) Available at http://www.infineon.com/customRF, Accessed November 2006. (Not BiCMOS; it's a competitive process.)

[6] Infineon. "PEB 3065 Signal Processing Subscriber Line Interface Codec Filter," Datasheet, http://www.infineon.com, Accessed November 2006.

[7] ITRS Working Group. *International Technology Roadmap For Semiconductors 2006 Update Overview And Working Group Summaries.* http://www.itrs.net, 2006, p.15.

[8] Ashok Bindra. "IEDM Charts The Course For Future-Generation Semiconductor Devices." *Electronic Design,* http://www.elecdesign.com/Articles/Print.cfm?ArticleID=5005, December 18, 2000.

[9] Don Tuite. "Mated To 0.13-µm CMOS, Foundry's SiGe npns Match RFCMOS For Less Cost." *Electronic Design,* http://www.elecdesign.com/Articles/Index.cfm?ArticleID=12936, July 6, 2006.

[10] L. E. Frenzel. "New Products For 3G Phones." *Electronic Design,* http://www.elecdesign.com/Articles/Print.cfm?ArticleID=4787, October 2, 2000.

[11] Ashok Bindra. "Analog/Mixed-Signal ICs: Introduction/Analog-To-Digital Converters." *Electronic Design,* http://www.elecdesign.com/Articles/Print.cfm?ArticleID=1424, January 7, 2002.

[12] Ashok Bindra. "Silicon-Germanium HBTs Merge With Mainstream CMOS Process." *Electronic Design,* http://www.elecdesign.com/Articles/Print.cfm?ArticleID=4522, July 10, 2000.

[13] Sigma Designs, Inc. "SMP8630 Secure Media Processor with Multiple A/V Codec." Preliminary Datasheet, http://www.sigmadesigns.com, January 2005.

[14] C. G. Fonstad. "Lecture 13 - Heterojunction Bipolar Transistors." Lecture slides, Course 6-772, Massachusetts Institute of Technology, http://ocw.mit.edu/OcwWeb/Electrical-Engineering-and-Computer-Science/index.htm, Spring 2003.

[15] M. K. Das, N. R. Das and P. K. Basu. "Performance Analysis of a SiGe/Si Heterojunction Bipolar Transistor for Different Ge-composition." *Proceedings of the XXVIIIth URSI General Assembly in New Delhi,* October 2005.

[16] Silvaco International. "SiGe/SiGeC Design Flow." Product brochure, Available at: http://www.silvaco.com/products/design_flows/sige_tech/sige.html, February 2006.

[17] D. L. Critchlow. "Recollections on MOSFET Scaling." *IEEE Solid-State Circuits Society News,* Vol. 12, pp. 19-22, Winter 2007.

[18] C. G. Fonstad. "Lecture 26 - The Current State of the Art." Lecture slides, Course 6-012, Massachusetts Institute of Technology, Available at: http://ocw.mit.edu/OcwWeb/Electrical-Engineering-and-Computer-Science/index.htm, Fall 2003.

[19] M. El-Gamal. "Introduction to the Special Issue on the 2003 IEEE Bipolar/BiCMOS Circuits and Technology Meeting." *IEEE Journal of Solid-State Circuits,* Vol. 39, pp. 1724-1725, October 2004.

[20] D. J. Walkey. "Introduction to the Special Issue on the IEEE Bipolar/BiCMOS Circuits and Technology Meeting." *IEEE Journal of Solid-State Circuits,* Vol. 40, pp. 1995-1996, October 2005.

[21] S. K. Reynolds, B. A. Floyd, U. R. Pfeiffer, T. Beukema, *et al.* "A Silicon 60-GHz Receiver and Transmitter Chipset for Broadband Communications." *IEEE Journal of Solid-State Circuits,* Vol. 41, pp. 2820-2831, December 2006.

[22] M. P. Flynn, P. K. T. Mok, Z. Wang W. A. Pribyl, *et al.* "Introduction to the Special Issue on the 2006 IEEE International Solid-State Circuits Conference." *IEEE Journal of Solid-State Circuits,* Vol. 41, pp. 2627-2631, December 2006.

[23] G. E. Zhang and M. M. Green. “A 10 Gb/s BiCMOS Adaptive Cable Equalizer.” *IEEE Journal of Solid-State Circuits,* Vol. 40, pp. 2132-2140, November 2005.

[24] H.-R. Lee, M.-S. Hwang, B.-J. Lee, Y.-D. Kim, *et al.* “A 1.2-V-Only 900-mW 10 Gb Ethernet Transceiver and XAUI Interface With Robust VCO Tuning Technique.” *IEEE Journal of Solid-State Circuits,* Vol. 40, pp. 2148-2158, November 2005. (Not BiCMOS. CMOS counterexample.)

[25] K. Nguyen, R. Adams, K. Sweetland, and H. Chen. “A 106-dB SNR Hybrid Oversampling Analog-to-Digital Converter for Digital Audio.” *IEEE Journal of Solid-State Circuits,* Vol. 40, pp. 2408-2415, December 2005. (Not BiCMOS. CMOS counterexample.)

[26] A. Ismail and A. A. Abidi. “A 3.1- to 8.2-GHz Zero-IF Receiver and Direct Frequency Synthesizer in 0.18-m SiGe BiCMOS for Mode-2 MB-OFDM UWB Communication.” *IEEE Journal of Solid-State Circuits,* Vol. 40, pp. 2573-2582, December 2005.

[27] A. J. Joseph, J. Dunn, G. Freeman, D. L. Harame, *et al.* “Product Applications and Technology Directions with SiGe BiCMOS.” *IEEE Journal of Solid-State Circuits,* Vol. 38, pp. 1471-1478, September 2003.

[28] A. J. Joseph, D. L. Harame, B. Jagannathan, D. Coolbaugh, *et al.* “Status and Direction of Communication Technologies—SiGe BiCMOS and RFCMOS.” *Proceedings of the IEEE,* Vol. 93, pp. 1539-1558, September 2005.

[29] B. Jagannathan, D. Greenberg, D. I. Sanderson, J.-S. Rieh, *et al.* “Speed and Power Performance Comparison of State-of-the-Art CMOS and SiGe RF Transistors.” *IEEE 2004 Topical Meeting on Silicon Monolithic Integrated Circuits in RF Systems,* pp. 115-118, 2004.

[30] C.Y. Chang, B.S. Wu, Y . K . Fang and R.H. Lee. “Amorphous Silicon Bipolar Transistor with High Gain (>12) and High Speed (>30ps).” *1985 International Electron Devices Meeting,* pp. 432-436, 1985. (Schottky emitter BJT as possible solution to sub-1V supply voltage.)

[31] J. L. Moll. “Comparison of Hot Electron and Related Amplifiers.” *IEEE Transactions on Electron Devices,* Vol. 10, pp. 299-304, September 1963. (Schottky emitter BJT as possible solution to sub-1V supply voltage.)

[32] R. J. Widlar. “Low Voltage Techniques.” *IEEE Journal of Solid-State Circuits,* Vol. SC-13, pp. 838-846, December 1978. (BJT operation at ~1V supply voltage. See [68].)

[33] K. Kimura. “Low Voltage Techniques for Bias Circuits.” *IEEE Transactions on Circuits and Systems,* Vol. 44, pp. 459-465, May 1997.

[34] R. J. Widlar. “New Developments in IC Voltage Regulators.” *IEEE Journal of Solid-State Circuits,* Vol. SC-6, pp. 2-7, February 1971. (Bandgap voltage reference.)

[35] J. Hodges, ed. “History of the Intel Microprocessor.” Website available at: http://www.i-probe.com/i-probe/ip_intel_4.html, December 1999.

[36] A. R. Alvarez, ed. *BiCMOS Technology and Applications.* 2nd ed. Boston: Kluwer, 1993.

[37] P. R. Gray, P. J. Hurst, S. H. Lewis and R. G. Meyer. *Analysis and Design of Analog Integrated Circuits.* 4th ed. New York: Wiley, 2001.

[38] B. Jagannathan, M. Khater, F. Pagette, J.-S. Rieh, *et al.* “Self-Aligned SiGe NPN Transistors with 285 GHz f_{MAX} and 207 GHz f_T in a Manufacturable Technology.” *IEEE Electron Device Letters,* Vol. 23, pp. 258-260, May 2002.

[39] M. E. Kim, A. K. Oki, G. M. Gorman, D. K. Umemoto, *et al.* “GaAs Heterojunction Bipolar Transistor Device and IC Technology for High-Performance Analog and Microwave Applications.” *IEEE Transactions on Microwave Theory and Techniques,* Vol. 37, pp. 1286-1303, September 1989.

[40] H. Kroemer. “Heterostructure Bipolar Transistors and Integrated Circuits.” *Proceedings of the IEEE,* Vol. 70, pp. 13-25, January 1982.

[41] C. D. Motchenbacher and J. A. Connelly. *Low-Noise Electronic System Design*. New York: Wiley, 1993, pp. 373, 376.

[42] Semiconductor Industry Association. “Worldwide Semiconductor Market Product Breakdown.” http://www.sia-online.org/downloads/Breakdown.pdf, 2005.

[43] R. J. Baker. *CMOS, Circuit Design, Layout and Simulation.* Piscataway: IEEE, 2005.

[44] Texas Instruments. “TLC220x, TLC220xA, TLC220xB, TLC220xY Advanced LinCMOS Low-Noise Precision Operational Amplifiers.” Datasheet, http://focus.ti.com/lit/ds/symlink/tlc2201.pdf, 1997.

[45] B. Sloman < bill.sloman@ieee.org>. “Re: steplike transistor noise.” Posting in Google newsgroup sci.electronics.design, February 1, 2001.

[46] A. B. Grebene. *Analog Integrated Circuit Design.* New York: Van Nostrand, 1972, p. 178.

[47] R. J. Baker. Website for [43]. http://cmosedu.com/cmos1/book.htm, 2007.

[48] C. Hull and S. V. Kishore. “BiCMOS Device Modeling for RF IC Design.” *IEEE BiCMOS Circuits and Technology Meeting,* pp. 17-22, 1999.

[49] I. Getreu. *Modeling the Bipolar Transistor.* Beaverton: Tektronix, Inc., 1976, p. 183.

[50] A. Ahmed. “Study of Low-Temperature Effects in Silicon-Germanium Heterojunction Bipolar Transistor Technology.” Masters Thesis, Georgia Institute of Technology, 2005.

[51] S. V. Huylenbroeck, S. Jenei, G. Carchon, A. Piontek, *et al.* “A 0.25µm SiGe BiCMOS Technology including Integrated RF Passive Components Optimised for Low Power Applications.” *33rd European Solid-State Device Research Conference,* http://www.imec.be/esscirc/essderc-esscirc-2003/papers/all/131.pdf, 2003.

[52] J. D. Cressler. “Assessing SiGe HBT Technology for Front-end Electronics Applications.” *5th International Meeting on Front-end Electronics,* 2003. (From slides at: http://www-ppd.fnal.gov/EEDOffice-w/conferences/FEE_2003/FEE2003talks/A3_Cressler.pdf.)

[53] S. Voldman, P. Juliana, N. Schmidt, A. Botula, *et al.* “ESD Robustness of a BiCMOS SiGe Technology.” *IEEE BiCMOS Circuits and Technology Meeting,* pp. 214-217, 2000.

[54] T. G. M. Kleinpenning. “Location of low-frequency noise sources in submicrometer bipolar transistors.” *IEEE Transactions on Electron Devices,* Vol. 39, pp. 1501-1506, 1992.

[55] J. Millman and C. C. Halkias. *Integrated Electronics: Analog and Digital Circuits and Systems.* New York: McGraw-Hill, 1972.

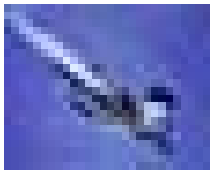
[56] A. S. Grove. *Physics and Technology of Semiconductor Devices.* New York: Wiley, 1967, p. 227.

[57] I. Lee and G. Kim. “Exponential Curvature-Compensated BiCMOS Bandgap References.” *IEEE Journal of Solid-State Circuits,* Vol. 29, pp. 1396-1403, November 1994.

[58] D. L. Harame, J. H. Comfort, J. D. Cressler, E. F. Crabbé, *et al.* “Si/SiGe Epitaxial-Base Transistors-Part I: Materials, Physics, and Circuits.” *IEEE Transactions on Electron Devices,* Vol. 42, pp. 455-468, March 1995.

[59] D. L. Harame, J. H. Comfort, J. D. Cressler, E. F. Crabbé, *et al.* “SiGe Epitaxial-Base Transistors-Part II: Process Integration and Analog Applications.” *IEEE Transactions on Electron Devices,* Vol. 42, pp. 469-482, March 1995.

[60] *The Strange Case of the Cosmic Rays*. 1957. Produced by Bell Telephone and directed by Frank Capra and William Hurtz. 55 min. Animated. Image Entertainment. DVD.



[61] H.C. de Graaff and J. G. de Groot. “The SIS Tunnel Emitter.” *1978 International Electron Devices Meeting,* Vol. 24, pp. 333-335, 1978.

[62] L. S. Vempati, J, D. Cressler, J. A. Babcock, R. C. Jaeger, *et al.* “Low-Frequency Noise in UHV/CVD Epitaxial Si and SiGe Bipolar Transistors.” *IEEE Journal of Solid-State Circuits,* Vol. 31, pp. 1458-1467, October 1996.

[63] A. I. Abdul-Rahim, C. D. Marsh, P. Ashbum and G. R. Booker. “Improved control of polysilicon emitter interfacial oxide using a UHV-compatible LPCVD cluster tool.” *IEEE 1997 Workshop on High Performance Electron Devices for Microwave and Optoelectronic Applications,* pp. 232-236, 1997.

[64] G. R. Wolstenholme, P. Ashburn, N. Jorgensen, D. Gold, *et al.* “Measurement and Modeling of the Emitter Resistance of Polysilicon Emitter Transistors.” *IEEE Bipolar Circuits and Technology Meeting,* pp. 55-58, 1998.

[65] T. H. Ning and D. D. Tang. “Method for Determining the Emitter and Base Series Resistances of Bipolar Transistors.” *IEEE Transactions on Electron Devices,* Vol. ED-31, pp. 409-412, April 1984.

[66] E. Sackinger and W. Guggenbuhl. “An analog trimming circuit based on a floating-gate device.” *IEEE Journal of Solid-State Circuits,* Vol. 23, pp. 1437-1440, December 1988.

[67] V. Srinivasan, G. J. Serrano, J. Gray and P. Hasler. “A Precision CMOS Amplifier Using Floating-Gate Transistors of Offset Cancellation.” *IEEE Journal of Solid-State Circuits,* Vol. 42, pp. 280-291.

[68] National Semiconductor Corp. “LM10 Operational Amplifier and Voltage Reference.” Datasheet, http://www.national.com/catalog, November 1994. (For [32].)

[69] J. S. Dunn, D. C. Ahlgren, D. D. Coolbaugh, N. B. Feilchenfeld, *et al.* “Foundation of RF CMOS and SiGe BiCMOS Technologies.” *IBM Journal of Research & Development,* Vol. 47, pp. 101-138, March/May 2003. (Available at: www.research.ibm.com/journal/rd/472/dunn.html)

[70] D. Hernandez-Garduno, and J. Silva-Martinez. “Continuous-Time Common-Mode Feedback for High-Speed Switched-Capacitor Networks.” *EEE Journal of Solid-State Circuits,* Vol. 40, pp. 1610-1617, August 2005.