Frequency Limits of InP-based Integrated Circuits

Mark Rodwell, E. Lind, Z. Griffith, S. R. Bank, A. M. Crook
U. Singisetti, M. Wistey, G. Burek, A.C. Gossard

Abstract—We examine the limits in scaling of InP-based bipolar and field effect transistors for increased device bandwidth. With InP-based HBTs, emitter and base contact resistivities and IC thermal resistance are the major limits to increased device bandwidth; devices with 1-1.5 THz simultaneous f₁ and fmax are feasible. Major challenges faced in developing either InGaAs HEMTs having THZ cutoff frequencies or InGaAs-channel MOSFETs having drive current consistent with the 22 nm ITRS objectives include the low two-dimensional effective density of states and the high bound state energies in narrow quantum wells.

I. INTRODUCTION

High -frequency integrated circuit technologies face relentless pressure from CMOS. 90 nm Silicon CMOS processes exhibit ~450 GHz power-gain cutoff frequencies (fmax) [1]. 45 nm processes, to be released into production this summer [2], will employ metal gates and high-K gate dielectrics, features which benefit both VLSI digital circuit performance and mm-wave amplification. Given the anticipated dates of introduction of 32 nm and 22 nm processes [2], it is likely that Si CMOS IC processes will provide transistors with >1 THz f₁ and fmax within the next four years. This places Si CMOS as a formidable competitive threat, not only to InP bipolar and field-effect transistors, but also to SiGe bipolar transistors.

Despite this rapid improvement in CMOS bandwidth, bipolar and InP processes retain significant advantages over Si MOS in analog/mixed-signal and microwave/mm-wave circuits. The high output conductance, limited gₑ, low breakdown voltage, and variable DC parameters of sub-90 nm Si MOSFETs makes precision mixed-signal design difficult [3]; bipolar transistors have high gₑ, low gₑ, high breakdown voltage, and precisely-controlled DC parameters, and consequently are attractive for precision mixed-signal ICs.

InP transistors obtain high bandwidths with less aggressive scaling and with simpler fabrication processes than are required for Si devices of comparable bandwidths. This is an advantage for low-volume fabrication of small-scale high-performance circuits. InP HBTs obtain significantly larger breakdown voltage that SiGe HBTs of a comparable bandwidth; their breakdown is much larger than that of highly scaled MOSFETs. This is an advantage for both mixed-signal ICs and mm-wave power amplifiers.

Consequently, despite the competitive pressure from CMOS, InP IC processes may survive in high-performance applications having low integration scales, much as the GaAs HBT remains the dominant device for cellular telephone power amplifiers. Further, InP HBTs appear to be far from their scaling limits, and cutoff frequencies beyond 1.5 THz appear to be feasible; this suggests new potential applications at sub-mm-wave frequencies.

Much broader markets may be found for InP-based electronics. Given the present understanding of the difficulties faced in scaling Si MOSFETs to < 22 nm gate length Lₑ, alternative channel materials --including InGaAs [4,5,6,7] -- are being considered for using in future MOS transistors in very-large scale ICs. The transport advantages of InGaAs channels --high mobilities and high carrier velocities-- are offset by several scaling difficulties associated with the low electron effective mass. Examining these scaling limits is relevant to the potential application of InGaAs to VLSI MOSFETs, and is equally relevant to the scaling potential of < 35 nm Lₑ InGaAs HEMTs.

II. INP BIPOLAR TRANSISTORS

InP HBTs have potential application in high-resolution ADCs and DACs with sampling rates in the range of 1-10 GS/s, in ~100 GHz gain-bandwidth product operational amplifiers for microwave signal processing, and in mm-wave and sub-mm-wave ICs, particularly power amplifiers. Present transistors [8] with 250 nm minimum feature size have attained 780 GHz power-gain cutoff frequencies (fmax) simultaneous with 424 GHz current-gain cutoff frequencies (f₁); devices at the 65 nm generation should attain cutoff frequencies well above 1 THz. This would enable e.g. digital ICs at ~450 GHz clock rate, 750 GHz monolithic power amplifiers, and compact ICs for sub-mm-wave frequency synthesis.

HBT bandwidths are increased by scaling [9,10,11]; both the scaling laws and a scaling roadmap are summarized in Table 1. It is worthwhile to review the underlying simplicity of these scaling laws. If we wish to improve by 2:1 the bandwidth of any circuit employing the HBT, we must reduce by 2:1 all transit delays and capacitances while retaining constant all resistances, all bias and signal voltages, and all bias and signal currents. The ratios of junction widths do not change with scaling (Figure 1); for brevity we omit the short departure required to justify this point. Thinning the
collector depletion layer by 2:1 will reduce the collector depletion-layer transit time by the required proportion but doubles junction capacitances per unit area; the collector and emitter junction areas must therefore also be reduced 4:1 so as to obtain the required 2:1 reduction in depletion capacitances. The emitter junction area has decreased 4:1; yet the emitter current and the emitter access resistivity must both remain constant. Both the emitter current density and the emitter specific access resistivity (resistance normalized to the contact area) must both be reduced 4:1. The 4:1 required increase in current density is consistent with the space-charge-limited current density (the Kirk effect) as this varies as the inverse square of depletion layer thickness.

In submicron InP HBTs, base access resistance is dominated by the Ohmic contact, the contact width is much less than the contact transfer length, and consequently the base resistance is approximately the base specific contact resistivity divided by the base Ohmic contact area. With a 4:1 reduction in junction areas, the base specific contact resistivity must thus be reduced 4:1.

HBT lithographic scaling laws are now primarily driven by thermal constraints. Approximating heat flow as half-cylindrical at radii \( r < L_E/2 \) and as hemispherical at greater distances the junction temperature rise of an isolated HBT on a thick substrate is

\[
\Delta T \approx \frac{P}{\pi K_{cbp} L_E} \ln \left( \frac{L_E}{W_E} \right) + \frac{1}{\pi K_{cbp} L_E}.
\]

(1)

\( K_{cbp} \) is the substrate thermal conductivity and \( P \) the dissipated power. We must reduce the emitter junction area \( L_E W_E \) by 4:1; maintaining constant emitter stripe length \( L_E \) while reducing the stripe width \( W_E \) by 4:1 results in only a moderate (logarithmic) increase in junction temperature with scaling.

We have now found the requirements for HBT scaling; the full scaling laws are shown in Table 1. The table shows not only transistor cutoff frequencies \( (f_t, f_{max}) \) but also key time constants (e.g. \( C_{cb} V_{logic} / I_c \) associated with digital gate delay [10]. Transistors for mm-wave amplification are designed for highest \( f_{max} \) and for \( f_t > f_{max} / 2 \), while HBTs for mixed-signal ICs are designed for minimum ECL gate delay [10]; transistors designed for highest feasible \( f_t \) but having \( f_{max} \ll f_t \) are of extremely limited utility in circuits.

Contact and thermal resistivities are presently the most serious barriers to scaling. Let us first consider thermal constraints. On large mixed-signal ICs, IC thermal resistance places perhaps the most severe scaling constraint. On an IC, transistor spacings \( D \) are small and must scale in inverse proportion to circuit bandwidth in order to scale wiring delays. Heat flow is then approximately half-cylindrical at radii \( r < L_E/2 \), hemispherical at radii \( L_E/2 < r < D/2 \), and planar at radii \( D/2 < r < T_{sub} \), where \( T_{sub} \) is the substrate thickness. The substrate temperature rise is then

\[
\Delta T_{sub} = \Delta T_{cylindrical} + \Delta T_{hemispherical} + \Delta T_{planar}
\]

\[
= \left( \frac{P}{\pi K_{cbp} L_E} \ln \left( \frac{L_E}{W_E} \right) + \frac{1}{\pi K_{cbp} L_E} \frac{1 - 1/D}{D} \right)
\]

(2)

In addition to the logarithmic temperature increase arising from narrow emitters, at fixed \( T_{sub} \) scaling causes \( \Delta T_{planar} \) to scale in proportion to the square of circuit bandwidth. The planar term can be reduced by extreme thinning of the substrate using wafer lapping or thermal vias [12].

Given a square IC of linear dimensions \( W_{chip} \) on a large copper heat sink, the package thermal resistance is approximated by planar and spherical regions, giving \( \Delta T_{package} = (1/2 + 1/\pi)(P_{chip} / K_{chip} W_{chip}) \). Because transistor

![Figure 1: HBT cross-section and critical dimensions. The emitter stripe extends a distance \( L_E \) perpendicular to the figure. \( W_E \) is the emitter junction width, \( W_{EB} \) the base-emitter sidewall spacer thickness, \( W_{cont} \) the width of the base Ohmic contact, and \( W_{under} \) the undercut of the collector junction under the base contacts.](image)
spacings vary as the inverse of IC bandwidth, $W_{ch}$ must vary by the same law, and $\Delta T_{package}$ therefore increases in direct proportion to increases in circuit bandwidth with scaling.

Using these thermal relationships, and scaling from an existing UCSB/Teledyne 150 GHz IC design the dissipation, wire lengths, transistor parameters, and circuit bandwidths, we can project (Figure 2) the resulting total junction temperature rise of a 2048-HBT CML digital integrated circuit as a function of digital clock rate. The substrate is thinned aggressively with the assumed use of thermal vias: $T_{sub} = 40 \mu m \cdot (150 \text{GHz} / f_{\text{clock}})$ to obtain these results.

IC design constraints are application specific, hence Figure 2 only illustrates $\Delta T$ calculation. It appears however that 450 GHz $f_{\text{clock}}$ is thermally feasible in 2000-transistor ICs, an integration scale typical of many ADCs and DACs. Higher clock rates than indicated in Figure 2 may be thermally feasible given improved circuit design. Present CML & ECL ICs must operate with a logic voltage swing $\Delta V_{\text{swing}} \approx 12 \cdot kT/q \approx 300 \text{ mV}$ in order to provide adequate voltage noise margin and operate with an interconnect characteristic impedance of $Z_o \approx 75 \Omega$. This stipulates a minimum switched current of $\Delta V_{\text{swing}}/Z_o = 4 \text{ mA per HBT}$. Low-voltage-swing-logic techniques, such as transimpedance input termination [13] can reduce the required HBT switched current and hence the IC dissipation.

Now let us consider scaling limits associated with contact resistivities. These must decrease in proportion to the inverse square of circuit bandwidth; 5 $\Omega\cdot\mu m^2$ base $\rho_b$ contact resistivity and 2 $\Omega\cdot\mu m^2$ emitter $\rho_e$ access resistivity are required for the 62 nm generation (440 GHz digital clock rate). In addition to the effects of doping and barrier potential, resistivity of $\text{ex-situ}$ deposited contacts is strongly influenced by surface oxides and cleaning procedures. For the base contacts, we observe the lowest resistivity with Pd solid-phase-reaction contacts [14], which penetrate oxides.

On recently reported HBTs [15,16], we measured 5 $\Omega\cdot\mu m^2$ for the emitter resistivity and $< 5 \Omega\cdot\mu m^2$ for the base resistivity. The emitter thus presents the more serious challenge. Therefore, we have recently investigated methods to reduce the emitter contact resistivity. Details will be reported elsewhere [17] ; several processes have been developed which provide between 0.5 and 1.0 $\Omega\cdot\mu m^2$ contact resistivity to N+ InGaAs layers. Electron degeneracy in the emitter-base depletion layer contributes an effective increase in the aggregate emitter resistance, proportional to $1/m^*_{\text{bulk}}$, of $-1.0 \Omega\cdot\mu m^2$.

Including the effects of emitter-base degeneracy, the emitter and base contact resistivities now achieved at UCSB are sufficient to meet the requirements for both the 125 nm and 65 nm scaling generations. Junction widths must be reduced and epitaxial layers thinned, but this simply requires efforts, albeit extensive, in development of the necessary device fabrication processes. Transistors with simultaneous 1.0 THz $f_1$ and 1.5 THz $f_{\text{max}}$ appear to be feasible. Such transistors would enable 450 GHz digital clock rates and 750 GHz power amplifiers. In the short term, our efforts now focus on realizing the 125 nm scaling generation device of Table 1. In the longer term, further efforts on base and emitter contact resistivity will likely have sufficient success to meet the requirements for the 31 nm scaling generation, a device $\sqrt{2}:1$ faster than the 62 nm device of Table 1. Note that the scaling requirements for sub-mm-wave amplification are somewhat less challenging than for high-speed digital ICs. In particular, integration scales are lower and hence IC thermal dissipation is a less serious challenge. Further, mm-wave power amplifiers can tolerate higher emitter access resistivities than ECL digital circuits; the scaling roadmap of Table 1 reflects digital circuit requirements for $R_e$.

![Figure 2: Calculated package and substrate temperatures rise, as a function of digital clock rate, for a 2048-HBT CML integrated circuit.](image)

Given recent interest in the use of III-V materials in future MOSFETs in VLSI [4,5,6,7], we examine several challenges faced in scaling these devices to meet the 22 nm ITRS performance goals. These scaling difficulties must also be addressed to enable development of > 1-THz-$f_1$. InGaAs Schottky-barrier field-effect transistors (HEMTs). Such devices would enable low-noise preamplifiers for sub-mm-wave radio receivers in the 50-500 GHz range.

Silicon MOSFETs were until recently mobility-limited, and significantly improved FET drive currents [18] were obtained through strain-induced increases in channel mobility. Investigation of high-mobility channel materials for VLSI was consequently motivated by the potential for further improved drive current ($I_d/W_e$) and decreased digital delay ($C_{\text{DS}}\Delta V/I_d$). Yet, the high mobility of InGaAs arises from a
low electron effective mass $m^*_{e}$, and low $m^*_{h}$ introduces numerous scaling difficulties. Further, short-channel InGaAs FETs have nearly-velocity-limited transconductance, and the benefit of high channel mobility is thus somewhat unclear. Finally, it has taken the past 17 years to double the bandwidth of InGaAs HEMTs [19, 20, 21], which suggests that serious scaling difficulties are already present.

As with bipolar transistor scaling, we seek a 2:1 reduction of and constant DC current, and constant DC voltages. While the well thickness $T_s$ should scale in proportion to $L_s$ so as to both proportionally scale $g_m/W_g$ and to maintain constant $g_{m}/g_{ds}$, thin wells present two difficulties. Thin wells reduce channel mobility [25] due to interface roughness scattering. Further, with thin wells the combined effects of the high ground state energy ($E_{sv} - h^2 \pi^2/2m^*_h$ in the infinite-width approximation) and high electron density ($E_s - E_{sv} \approx \frac{q^2 n}{c_{dos}}$) can naturally raise the channel Fermi energy to that of the L valley minimum, populating these bands and greatly reducing mobility. The low electron effective mass therefore constrains both vertical scaling and the maximum useful channel carrier density.

Fortunately, very large channel mobility is not required. Because $I_s \approx c_n v_{th} W_g (V_{gs} - V_{th})/L_s$, it is sufficient to maintain a mobility $\mu_s \gg \nu_{th} L_s/(V_s - V_{th})$. In practice, if $(V_s - V_{th}) \approx 0.7 \text{ V}, \nu_{th} = 3.3 \times 10^7 \text{ cm}/\text{s}$, and $L_s = 22 \text{ nm}$, $\mu_s = 1300 \text{ cm}^2/\text{V} \cdot \text{s}$ results in $I_s$ only 10% below that which would result from infinite mobility.

In our efforts to develop InGaAs-channel MOSFETs for the ITRS $L_s \approx 22$ nm generation, we seek $I_s/W_g > 3.5 \text{ mA}/\mu\text{m}$ at $(V_s - V_{th}) = 0.7 \text{ V}$ and consequently require $g_m/W_g > 5 \text{ mS}/\mu\text{m}$. Assuming $\nu_{th} = 3.3 \times 10^7 \text{ cm/s}$, this implies $N_s \approx 7 \times 10^{12}/\text{cm}^2$. Aspect ratio requirements stipulate that $T_s < 5 \text{ nm}$. Assuming infinitely deep wells and parabolic bands, the L valleys will be populated for $N_s > 6 \times 10^{17}/\text{cm}^2$; under more realistic calculations this may not occur until $N_s$ substantially exceeds $10^{19}/\text{cm}^2$.

Regarding these requirements, preliminary mobility data from growth experiments is encouraging. At the target $g_{m}/W_g$, $R/W_g$ should be below $\approx 20 \Omega \cdot \mu\text{m}$, well below that now achieved with InGaAs HEMTs [20]; we seek to achieve such low access resistivities using self-aligned N+ InGaAs regrown source-drain regions in combination with $< 1 \Omega \cdot \mu\text{m}^2$ resistivity Ohmic contacts similar to those now employed at UCSB for emitter contacts to HBTs.
ACKNOWLEDGMENTS

Ongoing development of InGaAs/InP-channel MOSFETs at UCSB is a collaboration with UCSD, Stanford, U. Minnesota, and U. Massachusetts, and we acknowledge the contributions of Professors Yuan Taur, Paul McIntyre, Susanne Stemmer, Chris Palmström, James Harris, Max Fischtz, Peter Asbeck, Andy Kummel, Chris Van de Walle, and of their students and postdoctoral researchers.

REFERENCES


[6] Yanning Sun, S.J. Koester, E.W. Kiewra, K.E. Fogel, D.K. Sadana, D.J. Webb, J. Fonpeyrine,J.-P. Locquet, M. Sousa, R. Germann, "Buried-channel In0.53Ga0.47As/In0.52Al0.48As MOS capacitors and transistors with HfO2 gate dielectrics" 2006 Device Research Conference, June, State College, PA, Page(s):49 - 50


