

**III-V MOSFETs: Scaling Laws, Scaling Limits, Fabrication Processes**


*Departments of ECE and Materials, University of California, Santa Barbara,*

*Now at the University of Notre Dame*

*Materials Science and Engineering Department, Stanford University,*

*Purdue University,*

*Departments of ECE and Materials, University of California, San Diego*

*University of Texas, Dallas,*

*University of Alabama, Huntsville*

**Abstract—III-V FETs are in development for both THz and VLSI applications.** In VLSI, high drive currents are sought at low gate drive voltages, while in THz circuits, high cutoff frequencies are required. In both cases, source and drain access resistivities must be decreased, and transconductance and drain current per unit gate width must be increased by reducing the gate dielectric thickness, reducing the inversion layer depth, and increasing the channel 2-DEG density of states. We here describe both nm self-aligned fabrication processes and channel designs to address these scaling limits.

**I. INTRODUCTION**

III-V transistors of ≈10 to 100 nm lithographic dimensions are being developed both for sub-mm-wave (0.3–3 THz) applications and for use in large-scale digital integrated circuits. Both applications demand improved transistor characteristics; both applications demand significant changes in the design and fabrication of the channel, of the source/drain access regions, and of the gate dielectric.

For application in VLSI, FET leakage currents must be low and drain drive current densities must be high despite low supply voltages. High intrinsic transconductance and low source/drain access resistivities are therefore required.

For application in THz ICs, high current-gain ($f_c$) and power-gain ($f_{max}$) cutoff frequencies are required. With present InGaAs HEMTs, $f_c$ is limited by parasitic capacitance charging times which are only reduced by increasing the FET transconductance per unit gate width. As with the VLSI application, the drive current and transconductance must be increased and the source access resistance reduced.

THz InGaAs HEMTs and InGaAs MOSFETs thus face several similar design challenges. To increase the transconductance of both HEMTs and MOSFETs, the gate barrier must be thinned, which increases gate leakage. In VLSI application, gate leakage must be very small, and an MOS structure with a wide-gap (insulating) gate dielectric is required. Even for HEMTs used in THz ICs, the wide-gap gate barrier semiconductor layer has been thinned to the point where gate leakage reduces microwave power gain; better barriers are needed. In both devices high transconductance implies both high carrier velocities and high carrier densities in the 2-dimensional electron gas. Semiconductors with low carrier effective mass provide high carrier velocities yet low 2-D densities of states hence low carrier densities, high effective mass provides low velocities yet high carrier densities. [1] This limitation must somehow be addressed. Both devices need low access resistances. Both devices need thin channels both for high transconductance and for low output conductance.

Design challenges with THz InGaAs HEMTs and InGaAs MOSFETs also differ in key aspects. Unlike THz HEMTs, where overall device dimensions can be much larger than the gate length, in VLSI the device packing density must be high hence all device dimensions must be small. In particular, in VLSI the source/drain contacts must have dimensions comparable to the gate length, placing greater demands on low-resistivity source/drain contacts. Similarly, while in THz HEMTs the N+ drain can have a large offset from the gate to reduce drain electrostatic coupling and consequently output conductance, in MOSFETs for VLSI both density and logic design requirements force the N+ drain region to be placed adjacent to or under the gate. Electrostatic design and vertical scaling of the VLSI device is therefore more demanding.

We describe below our efforts to develop III-V MOSFETs for VLSI. Although III-V MOS gate dielectrics [2, 3] remain an area of intense development, we focus here on device design and on development of process flows for fabrication of nm devices. Since their low 2-dimensional density of states makes III-V channel materials uncompetitive for application in nm FETs, we also discuss modified III-V channel designs which address this limitation.

**II. FET SCALING LAWS**

First consider FET scaling laws (Table 1) [4]. To increase bandwidth $\gamma$ : 1, capacitances and transit delays must be reduced $\gamma : 1$ while maintaining constant voltages, currents, and resistances. In InGaAs FETs with $L_e$ ~35 nm, the gate-source $C_{gs}$, $C_{gs} \propto \epsilon W_e$, and gate-drain $C_{gd}$, $C_{gd} \propto \epsilon W_g$, fringing capacitances are a substantial fraction of the total capacitance, and consequently limit $f_c$. $C_{gs}$ and $C_{gd}$ are only weakly dependent on lateral
Table 1: Constant-voltage / constant-velocity FET scaling laws: changes required for $\gamma : 1$ increased bandwidth in an arbitrary circuit

<table>
<thead>
<tr>
<th>parameter</th>
<th>law</th>
<th>parameter</th>
<th>law</th>
</tr>
</thead>
<tbody>
<tr>
<td>gate length $L_g$, source-drain contact lengths $L_{S/D}$ (nm)</td>
<td>$\gamma^{-2}$</td>
<td>electron density</td>
<td>$\gamma$</td>
</tr>
<tr>
<td>gate width $W_g$ (nm)</td>
<td>$\gamma^{-1}$</td>
<td>injection velocity (m/s)</td>
<td>$\gamma^0$</td>
</tr>
<tr>
<td>equivalent oxide thickness $T_{eq} = T_{ox} \cdot e_{ox} / e_{oxide}$ (nm)</td>
<td>$\gamma^{-1}$</td>
<td>drain current $I_d = qn \cdot V_{source}$ (mA)</td>
<td>$\gamma^0$</td>
</tr>
<tr>
<td>dielectric capacitance $C_{ox} = e_{ox} \cdot L_g \cdot W_g / T_{eq}$ (F)</td>
<td>$\gamma^{-2}$</td>
<td>gate-source, gate-drain fringing capacitances $C_{gs} \propto \varepsilon W_g$, $C_{gd} \propto \varepsilon W_g$ (F)</td>
<td>$\gamma^{-1}$</td>
</tr>
<tr>
<td>wavefunction mean depth $T_{mu}$ (nm)</td>
<td>$\gamma^{-2}$</td>
<td>S/D access resistances $R_{S}$, $R_{D}$ ($\Omega$)</td>
<td>$\gamma^0$</td>
</tr>
<tr>
<td>wavefunction depth capacitance $C_{depth} = e_{ox} \cdot L_g \cdot W_g / T_{mu}$ (F)</td>
<td>$\gamma^{-2}$</td>
<td>S/D contact resistivities $\rho_{S}$ ($\Omega - \mu m^2$)</td>
<td>$\gamma^2$</td>
</tr>
<tr>
<td>DOS capacitance (ballistic case) $C_{\text{dos}} = q^2 \cdot g(m \cdot m)^{1/2} \cdot L_g \cdot W_g / 2 \pi h^2$ (F)</td>
<td>$\gamma^{-2}$</td>
<td>temperature rise (one device, K) $W_g^{-1}$</td>
<td>$\gamma^1$</td>
</tr>
</tbody>
</table>

Consider drive current scaling in the ballistic limit. $I_d = qn \cdot v_{inf}$ is determined by the carrier injection velocity $v_{inf}$ and the sheet carrier concentration $n_s = (C_{g-a} / L_{W_g}) (V_{g-s} - V_{th}) / q$, where the gate-channel capacitance $C_{g-a} = (1 / C_{ox}) + (1 / C_{depth}) + (1 / C_{dos})$ is the series combination of dielectric $C_{ox} = e_{ox} \cdot L_g \cdot W_g / T_{eq}$, wavefunction depth capacitance $C_{depth} = e_{ox} \cdot L_g \cdot W_g / T_{mu}$, and density of states $C_{\text{dos}} = q \cdot d n / d E$ capacitances. $T_{mu}$ is here the wavefunction mean depth. In the ballistic case, $C_{\text{dos}} = q^2 \cdot g(m \cdot m)^{1/2} \cdot L_g \cdot W_g / 2 \pi h^2$, where $g$ is the number of populated valleys, and $m$ and $m$ the effective masses parallel and perpendicular to transport; near equilibrium, $C_{\text{dos}} = 2.1$ larger. Given ballistic transport [5] and assuming degenerate carrier concentrations, $E_g - E_{val} >> kT$, $v_{inf} = (4 / 3 \pi) (2 (E_g - E_{val}) / m) \cdot n_s = (4 / 3 \pi) (2qC_{g-a} (V_{g-s} - V_{th}) / m_{C_{ox}})^{1/2}$. We scale by maintaining constant $v_{inf}$ while reducing $C_{ox} / L_{W_g}$, $C_{\text{depth}} / L_g \cdot W_g$, and $C_{\text{dos}} / L_g \cdot W_g$ by $\gamma : 1$ so as to increase $n_s$ by $\gamma : 1$. This requires fixed transport mass $m$, $T_{eq}$, and $T_{mu}$ reduced $\gamma : 1$, and $C_{\text{dos}}$ increased $\gamma : 1$ by increasing $g$ of valleys or the perpendicular mass.

The FET is scaled such that the on-state current density $I_d / W_g$ (mA/\mu m) varies as $\gamma^{-3}$ while the current per unit source and drain Ohmic contact area (mA/\mu m$^2$) varies as $\gamma^{-1}$. It is well understood that difficulties in reducing $T_{eq}$ (gate leakage by tunneling) will impede constant-voltage FET scaling; note also that $T_{eq}$ must scale as $\gamma^{-3}$, requiring thinner wells or stronger confinement of the wavefunction in the well by strong vertical fields, and $(R_{S} + R_{D}) / W_g$ must scale as $\gamma^{-1}$, requiring a $\gamma : 1$ reduction in contact resistivity $\rho$, and increased carrier concentrations in the access regions. Design goals include low access resistance, high drive current density, thin wells, high sheet carrier density, and gate barriers that are both thin and high in energy.

To out-perform future scaled Si MOSFETs, drive currents must exceed 1-2 mA/\mu m at 300 mV gate overdrive $(V_{g-s} - V_{th})$. We must develop Ohmic contacts of $\sim 0.5 \Omega - \mu m^2$ contact resistivity; this resistivity must not increase when operating $\sim 150$ mA/\mu m$^2$ current density, nor can the contact metals diffuse under such high current and thermal stress through device junctions only $\sim 3-5$ nm below the surface. $T_{mu}$ must be at most 2-3 nm.

We describe our efforts to develop process to fabricate FETs having such parameters. We must also consider changes in the channel design necessary to enable the target high current densities.

III. DENSITY-OF-STATES LIMITS AND HIGH CURRENT DENSITY CHANNELS

We now examine the density-of-states limit to drive current and modified channel designs which address this limit.

Low transport mass produces high carrier velocities but low charge densities while high transport mass produces low carrier velocities but high charge densities. At a given dielectric thickness $T_{eq}$, there is an optimum $m$ maximizing $I_d$. We find
Table 2 Parameters of \( \Gamma, L, \) and X-valleys for several suitable semiconductors

<table>
<thead>
<tr>
<th>material</th>
<th>substrate</th>
<th>( \Gamma ) valley</th>
<th>( X ) valleys*</th>
<th>( L ) valleys</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{In}<em>{0.5}\text{Ga}</em>{0.5}\text{As} )</td>
<td>InP</td>
<td>( m^*/m_0 ) = 0.045</td>
<td>( m_\parallel/m_0 ) = 1.29</td>
<td>( m_{\perp}/m_0 ) = 0.19</td>
</tr>
<tr>
<td>InAs</td>
<td>InP</td>
<td>( m^*/m_0 ) = 0.026</td>
<td>( m_\parallel/m_0 ) = 1.13</td>
<td>( m_{\perp}/m_0 ) = 0.16</td>
</tr>
<tr>
<td>GaAs</td>
<td>GaAs</td>
<td>( m^*/m_0 ) = 0.067</td>
<td>( m_\parallel/m_0 ) = 1.3</td>
<td>( m_{\perp}/m_0 ) = 0.22</td>
</tr>
<tr>
<td>GaSb</td>
<td>GaSb</td>
<td>( m^*/m_0 ) = 0.039</td>
<td>( m_\parallel/m_0 ) = 1.51</td>
<td>( m_{\perp}/m_0 ) = 0.22</td>
</tr>
<tr>
<td>Si</td>
<td>Si</td>
<td>( m^*/m_0 ) = 0.92</td>
<td>( m_\parallel/m_0 ) = 0.19</td>
<td>( m_{\perp}/m_0 ) = 0.19</td>
</tr>
</tbody>
</table>

\( I_{\parallel}/W_k = J_0 \cdot K_1 \cdot (V_{g\parallel} - V_{th\parallel}) / (1V)^{3/2}, \) (1)

where

\[ J_0 = \left( \frac{4}{3\pi} \right) \left( \frac{2q}{m_0} \right)^{1/2} \left( \frac{q^2 m_0}{2}\right) \left( 1V \right)^{3/2} = 84 \text{ mA}/\mu\text{m} \] (2)

and

\[ K_1 = \frac{n \cdot (m_\parallel/m_{\perp})^{1/2}}{(1 + (C_{dlo} / C_{equiv}) \cdot g \cdot (m^2_{\parallel} m_{\perp}^{1/2} / m_{\perp}))^{3/2}} \] (3)

is the normalized current density. \( C_{equiv} = [1/C_{eq} + 1/C_{equiv}] \) is \( C_{depq} \) and \( C_{eq} \) in series, while \( C_{dlo} = q^2 m_{\parallel} L W_k / 2 \). Given one isotropic valley \( (m_\parallel = m_{\perp} = m^* \) or \( g = 1 \) and 1 \text{ nm} \text{ total equivalent dielectric thickness EOT (i.e. } C_{equiv} = e_{SiO_2} \cdot L W_k / (1 \text{ nm}) \text{), highest current is obtained for } m^*/m_{\perp} = 0.05, \text{ while for 0.3 \text{ nm} EOT, peak } I_{\parallel} \text{ is obtained at } m^*/m_{\perp} = 0.2; \text{ given one isotropic valley, low } m^* \text{ gives low } I_{\parallel} \text{ in nm FETs [6], though low } m^* \text{ reduces transit time for any EOT. Note than for Si } \{100\} \text{ FETs [6], } m^*/m_{\perp} = 0.19 \text{ and } g = 2. \)

Consider a 3 \text{ nm} \{100\} GaAs well with strained AlSb barriers. The L bound states lie 177 meV above that of \( \Gamma \). Equilibrium (not ballistic transport) analysis uses Schrödinger-Poisson, the effective mass approximation, and parabolic bands. 0.66 \text{ nm} Al\text{O}_2 and 0.34 nm AlSb lie between the well and gate, giving \( T_\parallel = 0.37 \text{ nm} \). Under strong inversion \( C_{dlo} / L W_k \geq 2.4 \mu F/cm^2 \), far below \( C_{eq} / L W_k = 9 \ \mu F/cm^2 \), and the high-mass L-valleys fill for \( n_\parallel > 2.4 \times 10^{11} \text{ cm}^{-2} \). Under ballistic transport, \( C_{dlo} \) and the maximum \( n_\parallel \) would both decrease 2:1.

Increased \( C_{dlo} \) and low \( m \) can be obtained by using L valley minima alone or combined with the \( \Gamma \) valley. The InGaAs, GaAs, and GaSb L-valleys [7] have low \( m_\parallel/m_\perp \) (0.062-0.1) and high \( m_\parallel/m_\perp \) (1.23-1.9). The L-valleys have \( \{111\} \) orientations, and transport in a (100) channel includes contributions from the high \( m_\parallel \). Using instead a \( \{111\} \) wafer, the L\{111\} valley is oriented vertically, and shows low transport masses \( (m_\parallel = m_{\perp} = m_\parallel) \) and high confinement mass \( (m_\parallel = m_\perp) \). The L\{111\}, \{111\}, and \{111\} minima show high transport mass [6] \((m_\parallel + 8m_\perp)/9 \) in one in-plane direction, but low confinement mass \((6m_\parallel m_\perp)/(m_\parallel + 8m_\perp) \). The X valleys have \( \{100\} \) orientations, in bulk InGaAs, GaAs, and GaSb have minima well above \( \Gamma \) and L, and in a \( \{111\} \) well have low \( m_\parallel = 3m_\parallel/(m_\parallel + 2m_\perp) \) quantization mass. In appropriate thin wells, the X and L\{111\}, \{111\}, and \{111\} quantized states are driven to high energies and depopulated. \( T_{\text{val}} \) can be selected to place \( \Gamma \) and L\{111\} at similar energies, doubling \( C_{dlo} \), or \( \Gamma \) driven in energy above L\{111\}, and transport provided in multiple L\{111\} valleys.

\( I_{\parallel}/W_k = (84 \text{ mA}/\mu\text{m}) \cdot K_1 \cdot (V_{g\parallel} - V_{th\parallel}) / (1V)^{3/2}, \) and \( g \) is the \# of valley minima.

Figure 1: \( \Gamma, L, \) and \( X \)-valley orientations for \( \{100\} \)- and \( \{111\} \)-oriented wafers

Figure 2: FET normalized drive current \( K_1 \) where \( I_{\parallel}/W_k = (84 \text{ mA}/\mu\text{m}) \cdot K_1 \cdot (V_{g\parallel} - V_{th\parallel}) / (1V)^{3/2} \), and \( g \) is the \# of valley minima.
Consider a 2.3 nm (111) GaAs well with strained AlSb barriers. $m_q$ is large, thus the first two \( L \) states are separated by only 84 meV. The $\Gamma$ state lies 41 meV above the lower \( L[111] \) state; 3 valleys are populated over a 300 mV range of $V_{gs}$. \( L[111] \) and [111] and X lie 175 and 288 meV above the lower \( L[111] \) state. In equilibrium simulation $n_s = 12 \times 10^{17}$ cm$^{-2}$ with $|V_{gs} - V_{th}| = 300$ mV; moderately higher $n_s$ does not populate heavy valleys. In inversion, $C_{e,\text{inv}}/L_s W_x \approx 4 \mu\text{F/cm}^2$. The benefit over the (100) design is larger in the ballistic case.

In InGaAs, GaAs, and GaSb, the \( L \)-valley $m_i$ is $>25:1$ larger than the $\Gamma$-valley mass, hence $T_{\text{inv}}$ can be made 5:1 smaller for a given quantization energy. $m_i$ is high in the barriers, hence multiple wells can be placed between $\sim 1$ nm barriers without significant well coupling hence energy redistribution. Multiple \( L[111] \) quantum wells can stacked to increase $g$ hence $C_{\text{inv}}$.

Consider a FET with two 0.66 nm (2 ML) (100) GaAs wells separated by strained 0.66 nm AlSb barriers. Given zero field, the two \( L[111] \) states split in energy by $< 40$ meV; for $|V_{gs} - V_{th}| = 300$ mV the separation is 56 meV. \( L[111] \), [111], and LX lie 322 and 346 meV above the lower \( L[111] \) state. The $\Gamma$ state is driven to high energy. In equilibrium, $n_s$ is driven to $12 \times 10^{18}$ cm$^{-2}$ with $|V_{gs} - V_{th}| \approx 300$ mV; moderately higher $n_s$ does not populate heavy valleys. $C_{e,\text{inv}}/L_s W_x \approx 4 \mu\text{F/cm}^2$. The advantage over $\Gamma \{100\}$ is greater for ballistic transport. A triple-well \( L[111] \) design gives similar results. In these FETs, the upper wells charge most strongly because of charge division between the wells' $C_{\text{inv}}$ and the well-well capacitance $C_{\text{well}} = \varepsilon L_s W_x / T_{\text{well}}$, where $T_{\text{well}}$ is the well pitch. With thin wells, and low $m_i$, $C_{\text{inv}}$ can be increased 1.5:1 to 2.2:1.
The designs above use very thin wells and barriers. It must be determined whether such layers can be grown and whether mobility is acceptable. The energy calculations must be refined. 2-4 ML GaSb and InAs wells [8,9] have been grown. Preliminary tightbinding calculations using an sp3d5s* basis [10] conducted for triple 1.1nm GaSb wells with 1.1nm AlSb barriers confirm the symmetry of the lowest state manifold and its expected transverse dispersion. Excited states are slightly lower than predicted by effective mass, but the design still appears viable. Experimental demonstration of such channel designs would enable III-V FETs to provide smaller carrier transit times and larger drive currents than Si MOSFETs even for gate dielectrics with equivalent thickness below 0.5 nm.

Figure 4: Process flow for III-V FETs with source/drain regrowth by MEE.

IV. FABRICATION PROCESSES FOR NM III-V MOSFETs
Established III-V HEMT structures do not well address scaling requirements of Section II. We have therefore developed a fully self-aligned InGaAs MOSFET process flow [11,12,13,14] (fig. 1). In this flow, 4.7 nm Al2O3 gate dielectric is deposited by ALD on a 5 nm In0.5Ga0.5As channel, the gate is formed by blanket W/Cr/SiO2 deposition and RIE etching, and thin ~25 nm Si N+ gate sidewalls formed. After etching the Al2O3, self-aligned S/D InAs N+ regions (50 nm thick, 8×10^{19} cm^{-3}, 23 Ω sheet resistance) are formed by migration enhanced epitaxy, and self-aligned S/D contacts formed by in-situ blanket evaporation of Mo (1-3Ω·μm contact resistance) and a subsequent height-selective etch [15]. Mesa isolation and back-end metal completes the process. Unlike HEMTs, no gate barrier is present in the S/D regions, the source and drain are fully self-aligned to the gate, and carrier densities in the S/D access regions are high (~1.5×10^{13} cm^{-3}). Figure 3 shows measured $I_d$ for a 200-nm-$L_g$ device.

Figure 5: Regrown S/D InGaAs FET, oblique view & cross-section

Figure 6: Common-source characteristics, 200 nm FET

ACKNOWLEDGMENTS
We acknowledge extensive discussions with H. Kroemer, B. Brar, J. Schulman, A. C. Gossard, and J. Albrecht. Devices were fabricated in the UCSB Nanofabrication Facility, a facility supported in part by the National Nanofabrication Users Network. This work made use of MRL Central Facilities supported by the MRSEC Program of the National Science Foundation under award No. MR05-20415
REFERENCES


Figure 7: Cross-section of regrown S/D InGaAs FET with a 27 nm gate