The Future of CMOS

David Pulfrey

CHRONOLOGY of the FET

• 1933 Lilienfeld’s patent (BG FET)
• 1965 Commercialization (Fairchild)
• 1991 “The most abundant object made by mankind” (C.T. Sah)
• 2003 The 10 nm FET (Intel)

90 Nanometer Technology (1 million units per week today)
SCALING: WHY DO IT?

- Increase speed
- Increase density
- Reduce cost (?)

SCALING: SPEED

![Graph showing gate delay vs. gate length for NMOS devices with different VDD and gate lengths. The graph includes data points for published data and Intel data.]
**SCALING: DENSITY**

- P4 (130): 55M in 146 mm²
- Itscott (90): 125M in 112 mm²
- Dathan (90): 140M in 87 mm²

**SCALING: COST**

- Nearly 7 Orders Of Magnitude Reduction in Cost/Transistor

Source: WSTS/Dataquest/Intel, 8/02
SCALING: FACTORY COST

Moore’s Law for Fabs!

THE SHRINKING FET

- $L_{\text{eff}}$ reduced 30X
- But devices are still “well tempered”

<table>
<thead>
<tr>
<th></th>
<th>CMOS 3</th>
<th>CMOS P18</th>
<th>CMOS P13</th>
<th>90NM</th>
</tr>
</thead>
<tbody>
<tr>
<td>L (nm)</td>
<td>3000</td>
<td>180</td>
<td>130</td>
<td>100</td>
</tr>
<tr>
<td>LINT (nm)</td>
<td>700</td>
<td>10</td>
<td>0</td>
<td>2.5</td>
</tr>
<tr>
<td>W (nm)</td>
<td>1000</td>
<td>160</td>
<td>190</td>
<td>150</td>
</tr>
<tr>
<td>TOX (nm)</td>
<td>85</td>
<td>4.1</td>
<td>2.8</td>
<td>2.3</td>
</tr>
<tr>
<td>NCH (cm$^{-3}$)</td>
<td>1.00E+16</td>
<td>3.90E+17</td>
<td>6.15E+17</td>
<td>8.37E+17</td>
</tr>
<tr>
<td>VDD (V)</td>
<td>5.0</td>
<td>1.8</td>
<td>1.2</td>
<td>1.0</td>
</tr>
<tr>
<td>VTHO (V)</td>
<td>0.95</td>
<td>0.47</td>
<td>0.35</td>
<td>0.24</td>
</tr>
</tbody>
</table>
THE SIGNIFICANCE OF $E_x$ AND $E_y$

- $E_x < E_y = \text{diode}$
- $E_x > E_y = \text{transistor}$

Well tempered means:
- keeping $E_x > E_y$
- and avoiding the short-channel effect

SHORT-CHANNEL EFFECT: $V_T$ depends on $L$

- Charge under gate due to $E$ from G, S and D
- Geometrical construction to estimate $V_T$ drop due to $E$ encroachment

$$\Delta V_T = \frac{\Delta Q_B}{C_{ox}}$$

- Reduce junction depth
Raised S and D

- Improves $I_{ON}$ by 20–30%

Current and $E_y$

<table>
<thead>
<tr>
<th>Year</th>
<th>VDD, V</th>
<th>L, nm</th>
<th>$E_y$, mV/nm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1985</td>
<td>5</td>
<td>3000</td>
<td>1.6</td>
</tr>
<tr>
<td>2003</td>
<td>1</td>
<td>100</td>
<td>10</td>
</tr>
<tr>
<td>2007</td>
<td>0.8</td>
<td>15</td>
<td>53</td>
</tr>
</tbody>
</table>

$$I_{Sat} = \frac{Z}{L} C_{ox} \mu \frac{(V_{GS} - V_T)^2}{2m}$$

$m = \text{body factor}$

$$I_{Sat} = C_{ox} \ Z \ \nu_{sat} (V_{GS} - V_T)$$

i.e. independent of L

and $f (V_{GS} - V_T)$
MOBILITY

- FETs don’t operate at high $E_y$ all the time, or over all of the channel.

- High mobility still very desirable to increase drive current

- Get high $\mu$ from strained-silicon channel

Si on SiGe: Tensile strain
Strained Si: breaking the symmetry

- 6 equivalent directions
- Intervalley scattering
- 2 sub-bands lowered in energy
- Reduced intervalley scattering
- Decreased effective mass (horizontal)

Strained Si: Relaxed sub-layers

IBM04
Decreasing \( (V_{DD} - V_T) \) means a loss of gate overdrive.

- Ultimate \( V_T \approx 0.2 \text{ V} \)
- Determined by \( I_{D,subt} \)
Control of \( I_{D,\text{subt}} \)

It’s done by capacitive control of the source-channel barrier height.

\[
V_{la} = V'_{S} = \frac{V_{GS}}{1 + \frac{C_S}{C_{ox}}} = \frac{V_{GS}}{m}
\]

\( m \): the body factor

\[
I_{D,\text{ideal}}(V_{GS} = 0) = I_{D,\text{thresh}} \exp \left[ -\frac{V_{t}}{mV_{t}} \right]
\]

This sets lower limit to \( V_T \), e.g., 0.2V.

\( I_{ON}/I_{OFF} \approx 10^4 \)

Sub-Threshold Slope

It’s the \( V_{GS} \) needed to reduce \( I_D \) by 10X.

\[
S = \left( \frac{d \log_{10} I_D}{dV_{GS}} \right)^{-1}
\]

\[
= m \cdot 2.303 \quad V_T = m \cdot 0.060 \quad V \text{ at } 300 \text{ K}
\]

\[
V_t = \frac{kT}{q} \quad \therefore \text{reduce } T
\]

Recall: \( m = 1 + \frac{C_B}{C_{ox}} = 1 + \frac{\varepsilon_s}{\varepsilon_{ox}} \cdot \frac{t_m}{W_d} \)

Need: small \( t_m \) and small \( N_d \)

Gate leakage

\( V_T \) compromise
Lecture 1

Cold-Electron Tunneling

de Broglie wavelength: \[ \lambda = \frac{h}{mv} = \frac{h}{\sqrt{2mKE}} \]

For an electron in Si at KE=\(\phi_{ox}/2\) : \(\lambda = 6.1 \text{ nm} \)

Electron could be either side of the barrier!

Tunneling Facts

Tunneling probability \(T = \left| \frac{A_{trans}}{A_{inc}} \right|^2 = \exp \left( -\frac{4\pi a}{\lambda} \right) \)

- For 180nm: 0.0002
- For 130nm: 0.0031
- For 90nm: 0.0088

What is a tolerable gate current?

![Tunneling Facts](image)
Ultimate Sub-Threshold Current

- S → D tunneling

- Expected to occur at L ≈ 10 nm

Transistor Off-state Leakage Trend

Research Data in Literature

Production Data
**Power constrained scaling limits**

<table>
<thead>
<tr>
<th>Device type</th>
<th>Application</th>
<th>$T$ (°C)</th>
<th>Power (W/cm²)</th>
<th>$V_{DD}$ (V)</th>
<th>$I_{DD}$ (mA/μm)</th>
<th>$V_{DD}$ (mV)</th>
<th>$I_{act}$ (μA/μm)</th>
<th>$I_{g}$ (μA/μm)</th>
<th>$I_{sub}$ (μA/μm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bulk</td>
<td>High performance</td>
<td>85</td>
<td>1000</td>
<td>0.8-1.2</td>
<td>5-100-2000</td>
<td>20</td>
<td>0.9-1.0</td>
<td>6-8.5</td>
<td>7-10</td>
</tr>
<tr>
<td>Bulk</td>
<td>Medium-high performance</td>
<td>85</td>
<td>10</td>
<td>0.6-1.0</td>
<td>50-40</td>
<td>270</td>
<td>1.2-1.4</td>
<td>8-11</td>
<td>10-21</td>
</tr>
<tr>
<td>Bulk</td>
<td>Moderate performance</td>
<td>85</td>
<td>1.0</td>
<td>0.6-1.0</td>
<td>6-4.5</td>
<td>360</td>
<td>1.4-1.6</td>
<td>9-42</td>
<td>19-24</td>
</tr>
<tr>
<td>Bulk</td>
<td>Low power</td>
<td>65</td>
<td>0.05</td>
<td>0.7-0.9</td>
<td>0.32-0.28</td>
<td>450</td>
<td>1.7-1.8</td>
<td>11-13</td>
<td>24-27</td>
</tr>
<tr>
<td>Bulk</td>
<td>Ultra-low power</td>
<td>40</td>
<td>&lt;0.001</td>
<td>0.7-1.0</td>
<td>&lt;0.0075</td>
<td>550-710</td>
<td>2.1-2.6</td>
<td>13-19</td>
<td>28-36</td>
</tr>
<tr>
<td>Bulk</td>
<td>Moderate-performance SRAM</td>
<td>85</td>
<td>5-1</td>
<td>0.9-1.2</td>
<td>60-10</td>
<td>240-310</td>
<td>1.3-1.5</td>
<td>10-13</td>
<td>20-26</td>
</tr>
<tr>
<td>Bulk</td>
<td>Low-power SRAM</td>
<td>65</td>
<td>0.1-0.01</td>
<td>0.9-1.2</td>
<td>1.5-1.15</td>
<td>300-470</td>
<td>1.6-2.0</td>
<td>12-16</td>
<td>25-22</td>
</tr>
<tr>
<td>Ultralow-power SRAM</td>
<td>40</td>
<td>0.0001</td>
<td>1.2</td>
<td>0.0018</td>
<td>590</td>
<td>2.4</td>
<td>20</td>
<td>59</td>
<td></td>
</tr>
</tbody>
</table>

It is power dissipation, rather than scaling, that will be the limiting factor. For example, can scale big servers more aggressively than portables and SRAMs.

High-k dielectrics

- High $T_{OX}$ needed to reduce gate leakage
- High $C_{OX}$ needed for $I_D$ and $S$
- Resolve conflict by increasing $\varepsilon$

\[
C_{ox} = \frac{\varepsilon_{ox}}{T_{ox}}
\]

<table>
<thead>
<tr>
<th>Dielectric</th>
<th>Dielectric constant (bulk)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Silicon dioxide ($SiO_2$)</td>
<td>3.9</td>
</tr>
<tr>
<td>Silicon nitride ($Si_3N_4$)</td>
<td>7</td>
</tr>
<tr>
<td>Aluminum oxide (Al$_2$O$_3$)</td>
<td>-10</td>
</tr>
<tr>
<td>Tantalum pentoxide (Ta$_2$O$_5$)</td>
<td>25</td>
</tr>
<tr>
<td>Lanthanum oxide (La$_2$O$_3$)</td>
<td>~21</td>
</tr>
<tr>
<td>Gadolinum oxide (Gd$_2$O$_3$)</td>
<td>~12</td>
</tr>
<tr>
<td>Yttrium oxide (Y$_2$O$_3$)</td>
<td>~15</td>
</tr>
<tr>
<td>Hafnium oxide (Hf$_2$O$_3$)</td>
<td>~20</td>
</tr>
<tr>
<td>Zirconium oxide (ZrO$_2$)</td>
<td>~23</td>
</tr>
</tbody>
</table>

High-k dielectrics: tunneling

Tunneling probability
\[
T = \frac{A_{max}}{A_{inc}} = \exp\left(-\frac{4\pi \phi}{A}\right)
\]

Tunneling probability
\[
T = \exp\left(-\frac{4\pi}{h} \int_0^\phi \sqrt{2m [V(x) - E]} \, dx\right)
\]

\[\therefore\] Need a high $\phi_{ox}$
High-k dielectrics: contenders

Also:
- must withstand poly activation (950°C)
- or use metal gate

Metal gate: self-alignment

Poly gates made self-alignment possible

Possibilities:
- Perhaps use sacrificial poly gate,
- then deposit metal.
- Co-evaporation of metals (Ti and Ni)
  to obtain different work functions,
- i.e., different $V_T$'s for NMOS and PMOS
  or for different blocks on same wafer.
**Metal gate and N\textsubscript{SUB}**

- If $V_T$ controlled by metal, perhaps can use undoped Si substrate.
- This would remove the problem of dopant fluctuations.

---

**Beyond Planar CMOS**

**Planar CMOS:**

- 10 nm prototypes demonstrated
- raised source and drain
- strained Si
- high $k$ dielectric
- metal gate
- limitation is power dissipation

**Further improvements:**

- Double gate CMOS
- SOI CMOS
Double-Gate CMOS

- Design flexibility - different $V_G$'s and $T_{ox}$'s
- SCE controlled by device geometry, not doping
- Can use undoped channel - reduces statistical fluctuations and Zener BD
- Increased $C_{ox}$ improves $I_{ON}$ and $S$

SCE: $V_T$ Roll-off

Note: benefit of shrinking $d_{Si}$
DG: Improved ON/OFF ratio

Recall: \( S = 2.303 \frac{m}{q} \frac{kT}{C_{ox}} \)  \( m = 1 + \frac{C_B}{C_{ox}} \)

DG doubles without reducing \( T_{ox} \)

Tends to zero (small \( d_{Si} \) and inversion from top and bottom)

- For same \( I_{OFF} \), set \( V_T \) 60mV lower, get more \( I_{ON} \)

DG example: FINFET

- DG is a deeply scalable FET, but fabrication is difficult

Wong02
### SOI CMOS

- More easily fabricated
- Ultra thin body
- No leakage through substrate
- Very low Cj
- Good device isolation for RF
- Technology of choice for SOC
- Not as deeply scalable as DG

[Diagram of SOI CMOS structure]

---

### SOI: state-of-the-art

- Small $L$, $x_f$, $d_{Si}$
- Raised S and D
- Fully depleted

[Scanning Electron Microscopy images of SOI structures showing $T_{Si} \leq L_g/3$]

- Silicide
- $L_g = 60$ nm
- $T_{Si} = 18$ nm
- Epi Raised S-D
- BOX
- Raised S-D using Selective Epi-Si Deposition

[Image credits: Gargini02a, Chau03]
Conclusion

- Planar CMOS: 10nm - THz operation - millions of transistors
- DG CMOS: reduced SCE - best sub threshold slope
  - high performance digital
- SOI CMOS: reduced leakage and parasitic C - RF capable
- Do we, or our children, need anything more?

References

Chau03 - ftp://download.intel.com/research/silicon/Chau%20DRC%20062303%20foils.pdf
Frank02 - Frank D.J. IBM J. R&D, v46, 235, 2002
Gargini02 - ftp://download.intel.com/research/silicon/PaoloISSUS0102.pdf
Gargini02a - ftp://download.intel.com/research/silicon/Paolo%20M2S2%200902.pdf
IBM03 - http://www-3.ibm.com/chips/services/foundry/offers/sige/5hp/
Morkoc04 - Morkoc H., WOCSDICE, Slovakia, 2004
Perlmutter04 - ftp://download.intel.com/research/silicon/Perlmutter053104.pdf
Raghavan00 - Raghavan G. et al., IEEE Spectrum, v37(10), 47, 2000
Rucker03 - Rucker R. et al., IEDM, paper 5.3, 2003
Wong02 - Wong H.S.P., IBM J. R&D, v46, 133, 2002
Yeo02 - Yeo K-S. et al., “CMOS/BiCMOS VLSI”, Prentice-Hall, 2002