Lecture #27

ANNOUNCEMENTS

• Tutebot demo contest Tues. May 11, 9:30-11am, 10 Evans

• Final Exam: Tues., May 18, 8-11am, 230 Hearst Gym; Exam is closed-book, except for 3 sheets (8.5 x 11 inches) of your notes. Exam is comprehensive.

OUTLINE

» Transistor scaling

» Silicon-on-Insulator technology

» Interconnect scaling

Reading (Rabaey et al.): Sections 3.5, 5.6
Transistor Scaling

Average minimum $L$ of MOSFETs vs. time

Steady advances in manufacturing technology (particularly lithography) have allowed for a steady reduction in transistor size.

$\sim 13\%$ reduction/year
(0.5$\times$ every 4-6 years)

How should transistor dimensions and supply voltage ($V_{DD}$) scale together?
Scenario #1: Constant-Field Scaling

- Voltages and MOSFET dimensions are scaled by the same factor $S > 1$, so that the electric field remains unchanged.
Impact of Constant-Field Scaling

(a) MOSFET gate capacitance:

\[ C'_{\text{gate}} = L' W' C'_{\text{ox}} = \left( \frac{L}{S} \right) \left( \frac{W}{S} \right) \cdot \left( \frac{\varepsilon_{ox}}{t_{ox}/S} \right) = \frac{C'_{\text{gate}}}{S} \]

(b) MOSFET drive current:

\[ I'_{\text{DSAT}} \propto C'_{\text{ox}} \frac{W'}{L'} \left( V'_{DD} - V'_{T} \right)^2 \approx \left( SC'_{\text{ox}} \right) \frac{W}{S} \left( \frac{V_{DD} - V_{T}}{S} \right)^2 \propto \frac{I'_{\text{DSAT}}}{S} \]

(c) Intrinsic gate delay:

\[ \frac{C'_{\text{gate}} V'_{DD}}{I'_{\text{DSAT}}} = \left( \frac{C_{\text{gate}}}{S} \right) \left( \frac{V_{DD}}{S} \right) = \left( \frac{C_{\text{gate}} V_{DD}}{I_{\text{DSAT}}} \right) \cdot \frac{1}{S} \]

✓ Circuit speed improves by S
Impact of Constant-Field Scaling (cont’d)

(d) Device density:

area required per transistor \( \propto WL' \)

\[ \begin{align*}
\text{\# of transistors per unit area} & \propto \frac{1}{W'L'} = \frac{1}{(W/S)(L/S)} = \frac{S^2}{WL} 
\end{align*} \]

(e) Power dissipated per device:

\[ P'_{\text{peak}} = I'_{\text{DSAT}} \cdot V'_{DD} = \left( \frac{I'_{\text{DSAT}}}{S} \right) \cdot \left( \frac{V_{DD}}{S} \right) = \frac{P_{\text{peak}}}{S^2} \]

(f) Power density:

\[ P'_{\text{peak}} \cdot \frac{1}{WL'} = \frac{P_{\text{peak}}}{S^2} \cdot \left( \frac{1}{(W/S)(L/S)} \right) = \frac{P_{\text{peak}}}{WL} \]

✓ Power consumed per function is reduced by \( S^2 \)
$V_T$ Scaling

- Low $V_T$ is desirable for high ON current:
  \[ I_{DSAT} \propto (V_{DD} - V_T)^\eta \quad 1 < \eta < 2 \]

- But high $V_T$ is needed for low OFF current:

$V_T$ cannot be aggressively scaled down!
Since $V_T$ cannot be scaled down aggressively, the power-supply voltage ($V_{DD}$) has not been scaled down in proportion to the MOSFET channel length:

<table>
<thead>
<tr>
<th>Feature Size ($\mu$m)</th>
<th>Power-Supply Voltage (V)</th>
<th>Gate Oxide Thickness (Å)</th>
<th>Oxide Field (MV/cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>5</td>
<td>350</td>
<td>1.4</td>
</tr>
<tr>
<td>1.2</td>
<td>5</td>
<td>250</td>
<td>2.0</td>
</tr>
<tr>
<td>0.8</td>
<td>5</td>
<td>180</td>
<td>2.8</td>
</tr>
<tr>
<td>0.5</td>
<td>3.3</td>
<td>120</td>
<td>2.8</td>
</tr>
<tr>
<td>0.35</td>
<td>3.3</td>
<td>100</td>
<td>3.3</td>
</tr>
<tr>
<td>0.25</td>
<td>2.5</td>
<td>70</td>
<td>3.6</td>
</tr>
</tbody>
</table>
Scenario #2: Generalized Scaling

- MOSFET dimensions are scaled by a factor $S > 1$; Voltages ($V_{DD}$ & $V_T$) are scaled by a factor $U > 1$

$$L' = L / S; \quad W' = W / S; \quad t'_{ox} = t_{ox} / S$$

$$V'_{DD} = V_{DD} / U$$

Note: $U$ is slightly smaller than $S$

(a) MOSFET drive current:

$$I'_{DSAT} \propto C'_{ox} \frac{W'}{L'} (V'_{DD} - V'_T)^2 \equiv (SC_{ox}) \left( \frac{W}{L/S} \right) \left( \frac{V_{DD} - V_T}{U} \right)^2 \propto \frac{SI'_{DSAT}}{U^2}$$

(b) Intrinsic gate delay:

$$\frac{C_{gate} V'_{DD}}{I'_{DSAT}} = \frac{C_{gate}/S (V_{DD}/U)}{(SI_{DSAT}/U^2)} = \left( \frac{C_{gate} V_{DD}}{I_{DSAT}} \right) \cdot \frac{U}{S^2}$$
Impact of Generalized Scaling

(c) Power dissipated per device:

\[ P'_{\text{peak}} = I'_{DSAT} \cdot V'_{DD} = \left( \frac{SI_{DSAT}}{U^2} \right) \cdot \left( \frac{V_{DD}}{U} \right) = \frac{SP_{\text{peak}}}{U^3} \]

(d) Power dissipated per unit area:

\[ P'_{\text{peak}} \cdot \frac{1}{W'L'} = \frac{SP_{\text{peak}}}{U^3} \cdot \left( \frac{1}{(W/S)(L/S)} \right) = \frac{S^3 P_{\text{peak}}}{U^3 WL} > \frac{P_{\text{peak}}}{WL} \]

- Reliability (due to high E-fields) and power density are issues!
Intrinsic Gate Delay ($C_{gate} V_{DD} / I_{DSAT}$)

![Graph showing the relationship between gate delay and gate length for NMOS transistors. The graph includes data points for published data and specific data points for Intel 30nm and Intel 20nm processes. The graph indicates that as gate length decreases, gate delay increases, with a noted threshold voltage $V_{DD} = 0.75V$.](image-url)
Transistors are fabricated in a thin single-crystal Si layer on top of an electrically insulating layer of SiO₂
- Simpler device isolation → savings in circuit layout area
- Low pn-junction & wire capacitances → faster circuit operation
- No "body effect"
- Higher cost
Interconnect Scaling

Relevant parameters:
- wire width $W$
- wire length $L$
- wire thickness $H$
- wire resistivity $\rho$
- wire-to-wire spacing $Z$
- inter-level dielectric (ILD)
  - thickness $t_{ILD}$
  - permittivity $\varepsilon_{ILD}$
For “local” (relatively short) interconnects:

- $W$, $Z$ and $t_{ILD}$ scale down by $S$
- $H$ is not scaled
  - avoids significantly increasing $R_{wire}$, but increases crosstalk
- $L$ scales down by a factor $S_L \leq S$

Wire capacitance scales by a factor $\varepsilon_c / S_L$, where $\varepsilon_c$ accounts for the impact of fringing & interwire capacitances.

For short & medium-length wires, the resistance of the driving logic gate dominates the wire resistance (i.e. $R_{dr} \gg R_{wire}$), so that the wire delay scales by $\varepsilon_c / S_L$. 
Global Interconnects

- For **global interconnects** (long wires used to route $V_{DD}$, GND, and voltage signals across a chip), the wire resistance dominates the resistance of the driving logic gate (i.e. $R_{\text{wire}} \gg R_{\text{dr}}$)

  $$R_{\text{wire}} C_{\text{wire}} \approx L^2$$

- The length of the longest wires on a chip increases slightly (~20%) with each new technology generation. In order to minimize increases in global interconnect delay, the cross-sectional area of global interconnects has not been scaled, i.e. $W$ and $H$ are not scaled down for global interconnects

  => Place global interconnects in separate planes of wiring
Interconnect Technology Trends

- Reduce the inter-layer dielectric permittivity
  - "low-k" dielectrics ($\varepsilon_r \approx 2$)
- Use more layers of wiring
  - average wire length is reduced
  - chip area is reduced

Intel 0.13µm Process (Cu)
Source: Intel Technical Journal 2Q02