Today

• VLSI Scaling Trends/Disciplines
• Effects
• Alternatives (cheating)

Scaling

• **Premise:** features scale "uniformly"
  – everything gets better in a predictable manner

• **Parameters:**
  • $\lambda$ (lambda) -- Mead and Conway (Day12)
  • $F$ -- Half pitch -- ITRS \((F=2\lambda)\)
  • $S$ -- scale factor -- Rabaey
  • $F'=SF$

ITRS Roadmap

• Semiconductor Industry rides this scaling curve
• Try to predict where industry going
  – (requirements…self fulfilling prophecy)

  • http://public.itrs.net

MOS Transistor Scaling (1974 to present)

$$S=0.7$$

[0.5x per 2 nodes]

Half Pitch (= Pitch/2) Definition
**Scaling Calculator +**

**Cycle Time:**

- $0.7x \rightarrow 0.7x$
- $250 \rightarrow 180 \rightarrow 130 \rightarrow 90 \rightarrow 65 \rightarrow 45 \rightarrow 32 \rightarrow 22 \rightarrow 16$
- $0.5x$

Node Cycle Time

- $N \rightarrow N+1 \rightarrow N+2$

*CARR(T) = Compound Annual Reduction Rate (B cycle time period, T)*

*CARR(T) = \((0.5)^{\frac{1}{2T \text{ yrs}}} - 1\)

*CARR(3 yrs) = -10.9%*

*CARR(2 yrs) = -15.9%*

---

**Scaling**

- Channel Length (L)
- Channel Width (W)
- Oxide Thickness ($T_{ox}$)
- Doping ($N_a$)
- Voltage (V)

---

**Full Scaling**

- Channel Length (L) $S$
- Channel Width (W) $S$
- Oxide Thickness ($T_{ox}$) $S$
- Doping ($N_a$) $1/S$
- Voltage (V) $S$

---

**Effects on Physical Properties?**

- Area
- Capacitance
- Resistance
- Threshold ($V_{th}$)
- Current ($I_d$)
- Gate Delay ($\tau_{gd}$)
- Wire Delay ($\tau_{wire}$)
- Power

- Go through full (ideal)
- …then come back and ask what still makes sense today.

---

**Area**

- $\lambda \rightarrow \lambda S$
- $A = L \times W$
- $A \rightarrow AS^2$

$S = 0.7$

[0.5x per 2 nodes]

---

**Capacity Scaling from Intel**
Capacitance

- Capacitance per unit area
  - $C_{ox} = \frac{\varepsilon_{SiO_2}}{T_{ox}}$
  - $T_{ox} = S \times T_{ox}$
  - $C_{ox} = C_{ox}/S$

Gate Capacitance

- $C_{gate} = A \times C_{ox}$
- $A = A \times S^2$
- $C_{ox} = C_{ox}/S$
- $C_{gate} = S \times C_{gate}$

Threshold Voltage

- $V_{TH} = S \times V_{TH}$

Current

- Saturation Current
  - $I_d = \left(\mu C_{ox}/2\right) (W/L) (V_{gs} - V_{TH})^2$
  - $V_{gs} = V \rightarrow S \times V$
  - $V_{TH} = S \times V_{TH}$
  - $W \rightarrow S \times W$
  - $L \rightarrow S \times L$
  - $C_{ox} = C_{ox}/S$
  - $I_d = S \times I_d$

Gate Delay

- $\tau_{gd} = Q/I = (C \times V)/I$
- $V = S \times V$
- $I_d \rightarrow S \times I_d$
- $C \rightarrow S \times C$
- $\tau_{gd} \rightarrow S \times \tau_{gd}$
Overall Scaling Results, Transistor Speed and Leakage. Preliminary Data from 2005 ITRS.

- **Intrinsic Transistor Delay**, \( \tau = CV/I \) (lower delay = higher speed)

- **Leakage Current**, \( \text{HP: standby power dissipation issues} \)

- **HP** → Target: 17%/yr, historical rate

- **LOP**

  - **LSTP** → Target: Isd,leak ~ 10 pA/um

- **HP = High-Performance Logic**

- **LOP = Low Operating Power Logic**

- **LSTP = Low Standby Power Logic**

- **17%/yr rate**

- **Planar Bulk MOSFETs**

  - **Advanced MOSFETs**

- **ITRS 2009 Transistor Speed**

Resistance

- \( R = \rho L/(W^*t) \)

- \( W \rightarrow S \times W \)

- \( L, t \) similar

- \( R \rightarrow R/S \)

Wire Delay

- \( \tau_{\text{wire}} = R \times C \)

- \( R \rightarrow R/S \)

- \( C \rightarrow S \times C \)

- \( \tau_{\text{wire}} \rightarrow \tau_{\text{wire}} \)

  - \( \ldots \)assuming (logical) wire lengths remain constant...

  - Assume short wire or buffered wire

  - Important cost shift we will have to watch

Power Dissipation (Dynamic)

- Capacitive (Dis) charging
  - \( P = (1/2)CV^2f \)

  - \( V \rightarrow S \times V \)

  - \( C \rightarrow S \times C \)

  - \( P \rightarrow S^3 \times P \)

  - Increase Frequency?

    - \( \tau_{\text{gd}} \rightarrow S \times \tau_{\text{gd}} \)

    - So: \( f \rightarrow f/S \) ?

    - \( P \rightarrow S^2 \times P \)

Effects?

- **Area** \( S^2 \)

- **Capacitance** \( S \)

- **Resistance** \( 1/S \)

- **Threshold (V\textsubscript{th})** \( S \)

- **Current (I\textsubscript{d})** \( S \)

- **Gate Delay (\( \tau_{\text{gd}} \))** \( S \)

- **Wire Delay (\( \tau_{\text{wire}} \))** \( 1 \)

- **Power** \( S^2 \rightarrow S^3 \)
### Power Density

- \( P \rightarrow S^2P \) (increase frequency)
- \( P \rightarrow S^3P \) (dynamic, same freq.)
- \( A \rightarrow S^2A \)

- Power Density: \( P/A \)
  - \( P/A \rightarrow P/A \) increase freq.
  - \( P/A \rightarrow S \times P/A \) same freq.

### Cheating…

- Don’t like some of the implications
  - High resistance wires
  - Higher capacitance
  - Atomic-scale dimensions
    - …. Quantum tunneling
  - Need for more wiring
  - Not scale speed fast enough

### Improving Resistance

- \( R = \rho L / (W \times t) \)
- \( W \rightarrow S \times W \)
- \( L, t \) similar
- \( R \rightarrow R/S \)

  - Don’t scale \( t \) quite as fast \( \rightarrow \) now taller than wide.
  - Decrease \( \rho \) (copper)

### Capacitance and Leakage

- Capacitance per unit area
  - \( C_{int} = f_{SOI} / T_{ox} \)
  - \( T_{int} = S \times T_{ox} \)
  - \( C_{int} = C_{int} / S \)

Reduce Dielectric Constant \( \varepsilon \) (interconnect)

and Increase Dielectric to substitute for scaling \( T_{ox} \)
(gate quantum tunneling)

---

#### ITRS 2009

<table>
<thead>
<tr>
<th>Year of Production</th>
<th>MPU/ASIC Metal 1 (M1)</th>
<th>Lg: Physical Lgate for High Performance logic (nm)</th>
<th>Lg: Physical Lgate for Low Operating Power (LOP) logic (nm)</th>
<th>EOT: Equivalent Oxide Thickness (nm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2016</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2017</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2018</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2019</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2020</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2021</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2022</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2023</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2024</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

#### Cross-section of Hierarchical Scaling—MPU Device

---

#### Cheating…

- Don’t like some of the implications
  - High resistance wires
  - Higher capacitance
  - Atomic-scale dimensions
    - …. Quantum tunneling
  - Need for more wiring
  - Not scale speed fast enough

---

#### Improving Resistance

- \( R = \rho L / (W \times t) \)
- \( W \rightarrow S \times W \)
- \( L, t \) similar
- \( R \rightarrow R/S \)

  - Don’t scale \( t \) quite as fast \( \rightarrow \) now taller than wide.
  - Decrease \( \rho \) (copper)
High-K dielectric Survey

<table>
<thead>
<tr>
<th>Dielectric</th>
<th>Dielectric constant (K)</th>
<th>Breakdown (V)</th>
<th>Conductivity (S/cm)</th>
<th>Leakage current (nA/cm)</th>
<th>Thermal stability (°C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Silicon dioxide</td>
<td>3.9</td>
<td>33</td>
<td>NA</td>
<td>&gt; 1000</td>
<td>&gt; 800 °C</td>
</tr>
<tr>
<td>Silicon nitride</td>
<td>7.7</td>
<td>21</td>
<td>2.8</td>
<td>&gt; 1000</td>
<td>&gt; 800 °C</td>
</tr>
<tr>
<td>Aluminum oxide</td>
<td>5.9</td>
<td>2.8</td>
<td>10</td>
<td>&gt; 1000</td>
<td>&gt; 800 °C</td>
</tr>
<tr>
<td>Titanium nitride</td>
<td>25</td>
<td>0.4</td>
<td>0.7</td>
<td>Not chemically stable</td>
<td></td>
</tr>
<tr>
<td>Lanthana oxide</td>
<td>30</td>
<td>0.7</td>
<td>2.5</td>
<td>Not chemically stable</td>
<td></td>
</tr>
<tr>
<td>Gallium arsenide</td>
<td>4.2</td>
<td>0.7</td>
<td>1.5</td>
<td>Not chemically stable</td>
<td></td>
</tr>
<tr>
<td>Yttrium oxide</td>
<td>15</td>
<td>2.3</td>
<td>1.5</td>
<td>Not chemically stable</td>
<td></td>
</tr>
<tr>
<td>Hafnium oxide</td>
<td>40</td>
<td>1.5</td>
<td>1.5</td>
<td>Not chemically stable</td>
<td></td>
</tr>
<tr>
<td>Zirconium oxide</td>
<td>25</td>
<td>1.4</td>
<td>1.5</td>
<td>Not chemically stable</td>
<td></td>
</tr>
<tr>
<td>Zirconium silicide</td>
<td>3</td>
<td>-0.1</td>
<td>-0.1</td>
<td>Not chemically stable</td>
<td></td>
</tr>
</tbody>
</table>

Intel NYT Announcement

- Intel Says Chips Will Run Faster, Using Less Power
  - NYT 1/27/07, John Markov
  - Claim: "most significant change in the materials used to manufacture silicon chips since Intel pioneered the modern integrated-circuit transistor more than four decades ago"
  - "Intel’s advance was in part in finding a new insulator composed of an alloy of hafnium... will replace the use of silicon dioxide."

Wire Layers = More Wiring

Typical chip cross-section illustrating hierarchical scaling methodology

...But

Power Dissipation (Dynamic)

- Capacitive (Dis) charging
  - \( P = \frac{1}{2} CV^2 f \)
  - \( V \rightarrow V \)
  - \( C \rightarrow S \times C \)
  - \( P \rightarrow S \times P \)
- Increase Frequency?
  - \( f \rightarrow f/S^2 ? \)
  - \( P \rightarrow P/S \)

If not scale V, power dissipation not scale down.
...And Power Density

- $P \rightarrow P/S$ (increase frequency)
- $P \rightarrow S \times P$ (same freq.)
- But... $A \rightarrow S^2 \times A$
- $P/A \rightarrow (1/S^3)P/A \quad \ldots \quad (1/S)P/A$

- Power Density Increases

...this is where some companies have gotten into trouble...

Intel on Power Challenge

[source: Borkar/Intel, Micro37, 12/04]

ITRS Vdd Scaling:
V Scaling more slowly than $F$

$V_{dd}$ vs Feature Size (nm)

CV$^2$ scaling from ITRS:
More slowly than $(1/F)^2$

Historical Power Scaling

Origin of Power Challenge

- Transistors per chip grow at Moore’s Law rate = $(1/F)^2$
- Energy/tr must decrease at this rate to keep constant
- $E/tr \propto CV^2f$

[Horowitz et al. / IEDM 2005]
Impact

- Can already place more transistors on a chip than we can afford to turn on.
- Power potential challenge/limiter for all future chips.

What Is A “Red Brick”?

- Red Brick = ITRS Technology Requirement with no known solution
- Alternate definition: Red Brick = something that REQUIRES billions of dollars in R&D investment

The “Red Brick Wall” - 2001 ITRS vs 1999

Table 1. 2001 Status of Red Brick Wall

<table>
<thead>
<tr>
<th>Year of Production</th>
<th>2001</th>
<th>2004</th>
<th>2005</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMAR drift length</td>
<td>120</td>
<td>120</td>
<td>120</td>
</tr>
<tr>
<td>Mobility</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>Short channel</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>Junction depth</td>
<td>1.2</td>
<td>1.2</td>
<td>1.2</td>
</tr>
<tr>
<td>Metal thickness</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Intermetal dielectric constant</td>
<td>1.2</td>
<td>1.2</td>
<td>1.2</td>
</tr>
</tbody>
</table>

Table 2. 1999 Status of Red Brick Wall

<table>
<thead>
<tr>
<th>Year of Production</th>
<th>1999</th>
<th>2002</th>
<th>2005</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMAR drift length</td>
<td>120</td>
<td>120</td>
<td>120</td>
</tr>
<tr>
<td>Mobility</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>Short channel</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>Junction depth</td>
<td>1.2</td>
<td>1.2</td>
<td>1.2</td>
</tr>
<tr>
<td>Metal thickness</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Intermetal dielectric constant</td>
<td>1.2</td>
<td>1.2</td>
<td>1.2</td>
</tr>
</tbody>
</table>

Conventional Scaling

- Ends in your lifetime
- ...perhaps in your first few years out of school...
- Perhaps already:
  - “Basically, this is the end of scaling.”
    - May 2005, Bernard Meyerson, V.P. and chief technologist for IBM's systems and technology group

ITRS 2009

- ARM Talk Tomorrow (Thursday)
  - Raisler 1:30pm – 3:00pm
- HW4 due Friday
Big Ideas
[MSB Ideas]

• Moderately predictable VLSI Scaling
  – unprecedented capacities/capability growth for engineered systems
  – change
  – be prepared to exploit
  – account for in comparing across time
  – …but not for much longer