Today

- VLSI Scaling Rules
- Effects
- Historical/predicted scaling
- Variations (cheating)
- Limits

Why Care?

- In this game, we must be able to predict the future
- Rapid technology advance
- Reason about changes and trends
- Re-evaluate prior solutions given technology at time X.

Why Care

- Cannot compare against what competitor does today
  - But what they can do at time you can ship
- Careful not to fall off curve
  - Lose out to someone who can stay on curve

Scaling

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

Feature Size

\( \lambda \) is half the minimum feature size in a VLSI process

[minimum feature usually channel width]
Scaling

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

Effects?

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

Area

- \lambda \rightarrow \lambda/\kappa
- A = L \times W
- A \rightarrow A/\kappa^2
- 130nm \rightarrow 90nm
- 50% area
- 2x capacity same area

Area Perspective

Capacity Scaling from Intel
Capacitance

- Capacitance per unit area
  \[ C_{ox} = \frac{\varepsilon_{SiO_2}}{T_{ox}} \]
  \[ T_{ox} \rightarrow T_{ox}/\kappa \]
  \[ C_{ox} \rightarrow \kappa C_{ox} \]

Threshold Voltage

Before:
\[ V_{th} = \frac{1}{\varepsilon_{SiO_2}} \left( -Q_{eff} + \frac{2Q_m V_{o} (V_{th} + V_{ss})^{1/2}}{V_{th} + \frac{V_{th}}{\kappa}} \right) + (W_f + \frac{V_{th}}{\kappa}) \]

Adjust \[ V_{bias} \rightarrow \frac{V_{th} + V_{ss}}{\kappa} \]

After:
\[ \frac{V_{th}}{\kappa} = \frac{1}{\varepsilon_{SiO_2}} \left( -Q_{eff} + \frac{2Q_m V_{o} (V_{th} + V_{ss})^{1/2}}{V_{th} + \frac{V_{th}}{\kappa}} \right) \]

Threshold Voltage

- \( V_{TH} \rightarrow V_{TH}/\kappa \)

Current

- Saturation Current
  \[ I_s = (\mu C_{ox}/2)(W/L)(V_{gs} - V_{th})^2 \]
  \[ V_{gs} \rightarrow V/\kappa \]
  \[ V_{th} \rightarrow V_{TH}/\kappa \]
  \[ W \rightarrow W/\kappa \]
  \[ L \rightarrow L/\kappa \]
  \[ C_{ox} \rightarrow \kappa C_{ox} \]
  \[ I_s \rightarrow I_s/\kappa \]
Gate Delay

\[ \tau_{gd} = \frac{Q}{I} = \frac{(CV)}{I} \]

- \( V \rightarrow V/\kappa \)
- \( I_d \rightarrow I_d/\kappa \)
- \( C \rightarrow C/\kappa \)
- \( \tau_{gd} \rightarrow \tau_{gd}/\kappa \)

Resistance

- \( R = \rho L / (W^*t) \)
- \( W \rightarrow W/\kappa \)
- \( L, t \) similar
- \( R \rightarrow \kappa R \)

Wire Delay

- \( \tau_{wire} = R \times C \)
- \( R \rightarrow \kappa R \)
- \( C \rightarrow C/\kappa \)
- \( \tau_{wire} \rightarrow \tau_{wire} \)

- …assuming (logical) wire lengths remain constant...
- Assume short wire or buffered wire
- (unbuffered wire ultimately scales as length squared)

Power Dissipation (Static Load)

- Resistive Power
  - \( P = V^*I \)
  - \( V \rightarrow V/\kappa \)
  - \( I_d \rightarrow I_d/\kappa \)
  - \( P \rightarrow P/\kappa^2 \)

Power Dissipation (Dynamic)

- Capacitive (Dis)charging
  - \( P = (1/2)CV^2f \)
  - \( V \rightarrow V/\kappa \)
  - \( C \rightarrow C/\kappa \)
  - \( P \rightarrow P/\kappa^3 \)

- Increase Frequency?
  - \( \tau_{gd} \rightarrow \tau_{gd}/\kappa \)
  - So: \( f \rightarrow \kappa f \) ?
  - \( P \rightarrow P/\kappa^2 \)
...and leakage

**The Leakage(s)...**

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

---

Intel on Leakage

**Projected Power (unconstrained)**

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

---

**Effects?**

- Area \( \frac{1}{\kappa^2} \)
- Capacitance \( \frac{1}{\kappa} \)
- Resistance \( \kappa \)
- Threshold (\( V_{th} \)) \( \frac{1}{\kappa} \)
- Current (\( I_d \)) \( \frac{1}{\kappa} \)
- Gate Delay (\( \tau_{gd} \)) \( \frac{1}{\kappa} \)
- Wire Delay (\( \tau_{wire} \)) 1
- Power \( \frac{1}{\kappa^2} \to \frac{1}{\kappa^3} \)

---

**ITRS Roadmap**

- Semiconductor Industry rides this scaling curve
- Try to predict where industry going
  \( \rightarrow \) (requirements...self fulfilling prophecy)
  - http://public.itrs.net

---

**MOS Transistor Scaling**

(1974 to present)

\[ S = 0.7 \]

[0.5x per 2 nodes]

[Source: 2001 ITRS - Exec. Summary, ORTC Figure]

[From Andrew Kahng]

---

**Half Pitch (= Pitch/2) Definition**

[Typical MPU/ASIC]

[Typical DRAM]

[Source: 2001 ITRS - Exec. Summary, ORTC Figure]

[From Andrew Kahng]
ITRS 2005

Figure 6 ITRS Product Technology Trends

What happens to delays?

- If delays in gates/switching?
- If delays in interconnect?
- Logical interconnect lengths?

Delays?

- If delays in gates/switching?
  - Delay reduce with $1/\kappa [\lambda]$

Delays

- Logical capacities growing
- Wirelengths?
  - No locality: $L \rightarrow \kappa$ (slower)
  - Rent's Rule
    - $L \rightarrow \eta(p<0.5)$
    - ($p > 0.5$)
Compute Density

- Density = compute / (Area * Time)
  - $\kappa^3$: compute density scaling
  - $\kappa^2$: gates dominate, $p < 0.5$
  - $\kappa^3$: moderate $p$, good fraction of gate delay
    - [p from Rent's Rule again – more on Day14]
  - $\kappa$: large $p$ (wires dominate area and delay)

Power Density

- $P \rightarrow P/\kappa^2$ (static, or increase frequency)
- $P \rightarrow P/\kappa^3$ (dynamic, same freq.)
- $A \rightarrow A/\kappa^2$
- $P/A \rightarrow P/A$ ... or ... $P/\kappa A$

Cheating...

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

Improving Resistance

- $R = \rho L/(W^*t)$
- $W \rightarrow W/\kappa$
- $L, t$ similar
- $R \rightarrow \kappa R$
  - Don't scale $t$ quite as fast.
  - Decrease $\rho$ (copper)

Capacitance and Leakage

- Capacitance per unit area
  - $C_{\text{ox}} = \epsilon_{\text{SiO}_2}/T_{\text{ox}}$
  - $T_{\text{ox}} \rightarrow T_{\text{ox}}/\kappa$
  - $C_{\text{ox}} \rightarrow \kappa C_{\text{ox}}$

  Reduce Dielectric Constant $\epsilon$ (interconnect) and Increase Dielectric to substitute for scaling $T_{\text{ox}}$ (gate quantum tunneling)
Threshold Voltage

Before:

\[ V_{th} = \frac{1}{C_{G0}} \left( -\phi + \left( 2\phi V_{th} + \phi + V_{th} \right)^{1/2} + \left( \phi + \phi \right) \right) \]

adjust \( V_{th} \) so \( \phi + V_{th} \) \( N \)

After:

\[ V_{th} = \frac{1}{C_{G0}} \left( -\phi + \left( 2\phi V_{th} + \phi + V_{th} \right)^{1/2} + \left( \phi + \phi \right) \right) \]

High-K dielectric Survey

Table 2: Various measured and predicted properties of high-k gate dielectrics. Data compiled from References [27,28,29,30,31,32,33,34].

<table>
<thead>
<tr>
<th>Dielectric</th>
<th>Various measured and predicted properties</th>
<th>Typical chip cross-section illustrating hierarchical scaling methodology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Silicon dioxide (SiO2)</td>
<td>9.0</td>
<td>9.6</td>
</tr>
<tr>
<td>Silicon nitride (Si3N4)</td>
<td>9.0</td>
<td>9.6</td>
</tr>
<tr>
<td>Aluminum oxide (Al2O3)</td>
<td>9.0</td>
<td>9.6</td>
</tr>
<tr>
<td>Titanium nitride (TiN)</td>
<td>9.0</td>
<td>9.6</td>
</tr>
<tr>
<td>Hafnium oxide (HfO2)</td>
<td>9.0</td>
<td>9.6</td>
</tr>
<tr>
<td>Tungsten silicide (WSi2)</td>
<td>9.0</td>
<td>9.6</td>
</tr>
<tr>
<td>Tungsten polycrystalline (WP)</td>
<td>9.0</td>
<td>9.6</td>
</tr>
<tr>
<td>Tantalum silicide (TaSi2)</td>
<td>9.0</td>
<td>9.6</td>
</tr>
<tr>
<td>Tantalum polycrystalline (TaP)</td>
<td>9.0</td>
<td>9.6</td>
</tr>
</tbody>
</table>

Intel Saturday 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.”

Typical chip cross-section illustrating hierarchical scaling methodology

[Unidentified]

Table 3: APDU Interconnect Technology Requirements—Near-term Trends (continued)

<table>
<thead>
<tr>
<th>Year of Production</th>
<th>2000</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cu thickness (nm)</td>
<td>8</td>
<td>10</td>
<td>12</td>
<td>14</td>
<td>16</td>
<td>18</td>
<td>20</td>
<td>22</td>
<td>24</td>
<td>26</td>
<td>28</td>
</tr>
<tr>
<td>Cu density (Joules/Unit)</td>
<td>4.5</td>
<td>4.5</td>
<td>4.5</td>
<td>4.5</td>
<td>4.5</td>
<td>4.5</td>
<td>4.5</td>
<td>4.5</td>
<td>4.5</td>
<td>4.5</td>
<td>4.5</td>
</tr>
</tbody>
</table>

Typical chip cross-section illustrating hierarchical scaling methodology
Improving Gate Delay

- \( \tau_{gd} = Q/I = (CV)/I \)
- \( V \rightarrow V/\kappa \)
- \( I_d = (\kappa C_{gd}/2)(W/L)(V_{gs}-V_{th})^2 \)
- \( C \rightarrow C/\kappa \)
- \( \tau_{gd} \rightarrow \tau_{gd}/\kappa \)
- Lower C.
- Don’t scale V.

\[ \frac{\tau_{gd}}{\kappa} \]

...But Power Dissipation (Dynamic)

- Capacitive (Dis)charging
  - \( P = (1/2)CV^2f \)
- \( V \rightarrow V/\kappa \)
- \( C \rightarrow C/\kappa \)
- \( P \rightarrow P/\kappa^3 \)

If not scale V, power dissipation not scale.

...And Power Density

- \( P \rightarrow P/\kappa \) (increase frequency)
- \( P \rightarrow P/\kappa \) (dynamic, same freq.)
- \( \Lambda \rightarrow \Lambda/\kappa^2 \)
- \( P/A \rightarrow \kappa P/A \) ... or ... \( \kappa^2 P/A \)

Power Density Increases
...this is where some companies have gotten into trouble...

Intel on Leakage

Physical Limits

- Doping?
- Features?

Physical Limits

- Depended on
  - bulk effects
    - doping
    - current (many electrons)
    - mean free path in conductor
  - localized to conductors
- Eventually
  - single electrons, atoms
  - distances close enough to allow tunneling
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

![Diagram of Dopants/Transistor](image1)

![Diagram of Electric Field and Potential](image2)

![Diagram of Technology Nodes](image3)

Electrons

- $e = 1.6 \times 10^{-19}$ C

- How many electrons?

![Diagram of Red Brick Wall - 2001 ITRS vs 1999](image4)

The “Red Brick Wall” - 2001 ITRS vs 1999

ITRS 2005 ...

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
Finishing Up...

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

Big Ideas
[MSB-1 Ideas]

• Uniform scaling reasonably accurate for past couple of decades
• Area increase $\kappa^2$
  – Real capacity maybe a little less?
• Gate delay decreases ($1/$)
• Wire delay not decrease, maybe increase
• Overall delay decrease less than ($1/$)