EE141- Spring 2003
Lecture 9
Technology Scaling

Announcements

- Midterm 1 Next Tu (6:30-8 pm in 277 Cory Hall)
  » Material: Everything up to today’s lecture.
  This is: Chapter 1, 3 and 5 (as well as part of Chapter 2)
- Project 1 will be launched next Th.
  Topic: Driver design
Today’s lecture

- Inverter chain sizing
- Technology Scaling

Sizing the Inverter Chain

\[ t_p = t_{p1} + t_{p2} + \ldots + t_{pN} \]

\[ t_{pj} \sim R_{unit} C_{unit} \left( 1 + \frac{C_{gin,j+1}}{C_{gin,j}} \right) \]

\[ t_p = \sum_{j=1}^{N} t_{p,j} = \sum_{j=1}^{N} \left( \sum_{i=1}^{C_{gin,N+1}} \frac{1 + \frac{C_{gin,j+1}}{C_{gin,j}}}{C_{gin,j}} \right), \quad C_{gin,N+1} = C_L \]
Optimum Delay and Number of Stages

When each stage is sized by $f$ and has same eff. fanout $f$:

$$f^N = F = \frac{C_L}{C_{gin,1}}$$

Effective fanout of each stage:

$$f = \sqrt[N]{F}$$

Minimum path delay

$$t_p = Nt_{p0} \left(1 + \sqrt[N]{F} / \gamma \right)$$

---

Optimum Number of Stages

For a given load, $C_L$ and given input capacitance $C_{in}$

Find optimal sizing $f$

$$C_L = F \cdot C_{in} = f^N C_{in} \quad \text{with} \quad N = \frac{\ln F}{\ln f}$$

$$t_p = Nt_{p0} \left(F^{1/N} / \gamma + 1 \right) = \frac{t_{p0} \ln F}{\gamma} \left(\frac{f}{\ln f} + \frac{\gamma}{\ln f} \right)$$

$$\frac{\partial t_p}{\partial f} = \frac{t_{p0} \ln F}{\gamma} \cdot \frac{\ln f - 1 - \gamma/f}{\ln^2 f} = 0$$

For $\gamma = 0$, $f = e$, $N = \ln F$  

$$f = \exp(1 + \gamma/f)$$
Optimum Effective Fanout $f$

Optimum $f$ for given process defined by $\gamma$

$$f = \exp\left(1 + \frac{\gamma}{f}\right)$$

$$f_{opt} = 3.6$$

for $\gamma = 1$

Normalized delay function of $F$

$$t_p = N t_{p0} \left(1 + \sqrt[N]{F} / \gamma \right)$$

<table>
<thead>
<tr>
<th>$F$</th>
<th>Unbuffered</th>
<th>Two Stage</th>
<th>Inverter Chain</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>11</td>
<td>8.3</td>
<td>8.3</td>
</tr>
<tr>
<td>100</td>
<td>101</td>
<td>22</td>
<td>16.5</td>
</tr>
<tr>
<td>1000</td>
<td>1001</td>
<td>6.5</td>
<td>24.8</td>
</tr>
<tr>
<td>10,000</td>
<td>10,001</td>
<td>202</td>
<td>33.1</td>
</tr>
</tbody>
</table>
### Buffer Design

<table>
<thead>
<tr>
<th>N</th>
<th>f</th>
<th>$t_p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>64</td>
<td>65</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>18</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td>4</td>
<td>2.8</td>
<td>15.3</td>
</tr>
</tbody>
</table>

### What about power consumption (and area)?

$$C_{tot} = C_i + eC_i + ... + e^N C_i$$

$$= C_i \left(1 + e + ... + e^N\right)$$

$$= C_i + C_i e^N + C_i e^N \left(1 + e + ... + e^{N-2}\right)$$

Overhead!! $(E^N - 1)/(E - 1)$

e.g. $C_L = 20\mu F; C_i = 50\mu F \Rightarrow N = 6$

Fixed: 20pF

Overhead: 11.66pF !!!
Transistor Sizing for Minimum Energy

- Goal: Minimize Energy of whole circuit
  - Design parameters: $f$ and $V_{DD}$
  - $t_p \leq t_{pref}$ of circuit with $f=1$ and $V_{DD}=V_{ref}$

$$t_p = t_{p0} \left( \left[ 1 + \frac{f}{\gamma} \right] + \left[ 1 + \frac{F}{f\gamma} \right] \right)$$

$$t_{p0} \propto \frac{V_{DD}}{V_{DD} - V_{TE}}$$
Transistor Sizing (2)

- Performance Constraint ($\eta=1$)

$$\frac{t_p}{t_{p,ref}} = \frac{t_{p,0}}{t_{p,0,ref}} \left(1 + \frac{2 + \frac{F}{f}}{3 + F}\right) = \frac{V_{DD}}{V_{ref}} \frac{V_{ref} - V_{TE}}{V_{DD} - V_{TE}} \left(1 + \frac{2 + \frac{F}{f}}{3 + F}\right) = 1$$

- Energy for single Transition

$$E = V_{DD}^2 C_{g1} \left[1 + \gamma(1 + f) + F\right]$$

$$\frac{E}{E_{ref}} = \left(\frac{V_{DD}}{V_{ref}}\right)^2 \left(\frac{2 + 2 f + F}{4 + F}\right)$$
Impact of Technology Scaling

Goals of Technology Scaling

- Make things cheaper:
  - Want to sell more functions (transistors) per chip for the same money
  - Build same products cheaper, sell the same part for less money
  - Price of a transistor has to be reduced
- But also want to be faster, smaller, lower power
Technology Scaling

- Goals of scaling the dimensions by 30%:
  - Reduce gate delay by 30% (increase operating frequency by 43%)
  - Double transistor density
  - Reduce energy per transition by 65% (50% power savings @ 43% increase in frequency)
- Die size used to increase by 14% per generation
- Technology generation spans 2-3 years

Technology Generations
## Technology Evolution (2000 data)

### International Technology Roadmap for Semiconductors

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology node [nm]</td>
<td>180</td>
<td>130</td>
<td>90</td>
<td>60</td>
<td>40</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>Supply [V]</td>
<td>1.5-1.8</td>
<td>1.5-1.8</td>
<td>1.2-1.5</td>
<td>0.9-1.2</td>
<td>0.6-0.9</td>
<td>0.5-0.6</td>
<td>0.3-0.6</td>
</tr>
<tr>
<td>Wiring levels</td>
<td>6-7</td>
<td>6-7</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>9-10</td>
<td>10</td>
</tr>
<tr>
<td>Max frequency [GHz], Local-Global</td>
<td>1.2</td>
<td>1.6-1.4</td>
<td>2.1-1.6</td>
<td>3.5-2</td>
<td>7.1-2.5</td>
<td>11-3</td>
<td>14.9-3.6</td>
</tr>
<tr>
<td>Max µP power [W]</td>
<td>90</td>
<td>106</td>
<td>130</td>
<td>160</td>
<td>171</td>
<td>177</td>
<td>186</td>
</tr>
<tr>
<td>Bat. power [W]</td>
<td>1.4</td>
<td>1.7</td>
<td>2.0</td>
<td>2.4</td>
<td>2.1</td>
<td>2.3</td>
<td>2.5</td>
</tr>
</tbody>
</table>


---

## Technology Evolution (1999)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Channel length (µm)</td>
<td>0.4</td>
<td>0.3</td>
<td>0.25</td>
<td>0.18</td>
<td>0.13</td>
<td>0.1</td>
</tr>
<tr>
<td>Gate oxide (nm)</td>
<td>12</td>
<td>7</td>
<td>6</td>
<td>4.5</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>$V_{DD}$ (V)</td>
<td>3.3</td>
<td>2.2</td>
<td>2.2</td>
<td>1.5</td>
<td>1.5</td>
<td>1.5</td>
</tr>
<tr>
<td>$V_T$ (V)</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
<td>0.6</td>
<td>0.6</td>
<td>0.6</td>
</tr>
<tr>
<td>NMOS $I_{Dsat}$ (mA/µm) (@ $V_{GE} = V_{DD}$)</td>
<td>0.35</td>
<td>0.27</td>
<td>0.31</td>
<td>0.21</td>
<td>0.29</td>
<td>0.33</td>
</tr>
<tr>
<td>PMOS $I_{Dsat}$ (mA/µm) (@ $V_{GE} = V_{DD}$)</td>
<td>0.16</td>
<td>0.11</td>
<td>0.14</td>
<td>0.09</td>
<td>0.13</td>
<td>0.16</td>
</tr>
</tbody>
</table>
ITRS Technology Roadmap Acceleration Continues

Technology Scaling (1)
Technology Scaling (2)

Number of components per chip

Technology Scaling (3)

$\tau_p$ decreases by 13%/year
50% every 5 years!

Propagation Delay
Technology Scaling (4)

(a) Power dissipation vs. year.

(b) Power density vs. scaling factor.

From Kuroda

Technology Scaling Models

- **Full Scaling (Constant Electrical Field)**
  
  ideal model — dimensions and voltage scale together by the same factor $S$

- **Fixed Voltage Scaling**
  
  most common model until recently — only dimensions scale, voltages remain constant

- **General Scaling**
  
  most realistic for today’s situation — voltages and dimensions scale with different factors
Scaling Relationships for Long Channel Devices

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Relation</th>
<th>Full Scaling</th>
<th>General Scaling</th>
<th>Fixed Voltage Scaling</th>
</tr>
</thead>
<tbody>
<tr>
<td>( W, L, L_nx )</td>
<td>( W/L )</td>
<td>1/S</td>
<td>1/S</td>
<td>1/S</td>
</tr>
<tr>
<td>( V_{DDP}, V_T )</td>
<td>( V/W_{dep} )</td>
<td>S</td>
<td>S^3/U</td>
<td>S^3</td>
</tr>
<tr>
<td>( N_{SUB} )</td>
<td>( V/W_{dep}^2 )</td>
<td>S</td>
<td>S^3/U</td>
<td>S^3</td>
</tr>
<tr>
<td>Area/Device</td>
<td>WL</td>
<td>1/S^2</td>
<td>1/S^2</td>
<td>1/S^2</td>
</tr>
<tr>
<td>( C_{ox} )</td>
<td>( 1/L_{ox} )</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>( C_L )</td>
<td>( C_{ox}WL )</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>( b_m, b_p )</td>
<td>( C_{ox}V/L )</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>( I_{av} )</td>
<td>( b_{n, p} V^2 )</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>( I_p ) (intrinsic)</td>
<td>( C_L V / I_{av} )</td>
<td>1/S</td>
<td>1/S U^2</td>
<td>1/S^2</td>
</tr>
<tr>
<td>( P_{av} )</td>
<td>( C_L V^2 / I_p )</td>
<td>1/S^2</td>
<td>S/U^3</td>
<td>S</td>
</tr>
<tr>
<td>PDP</td>
<td>( C_L V^2 )</td>
<td>1/S^3</td>
<td>1/S U^3</td>
<td>1/S</td>
</tr>
</tbody>
</table>

Table 3.1: Scaling Relationships for Long Channel Devices

Transistor Scaling
(velocity-saturated devices)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Relation</th>
<th>Full Scaling</th>
<th>General Scaling</th>
<th>Fixed-Voltage Scaling</th>
</tr>
</thead>
<tbody>
<tr>
<td>( W/L, L_n )</td>
<td>( W/L )</td>
<td>1/S</td>
<td>1/S</td>
<td>1/S</td>
</tr>
<tr>
<td>( V_{DDP}, V_T )</td>
<td>( V/W_{dep} )</td>
<td>S</td>
<td>S^3/U</td>
<td>S^3</td>
</tr>
<tr>
<td>Area/Device</td>
<td>WL</td>
<td>1/S^2</td>
<td>1/S^2</td>
<td>1/S^2</td>
</tr>
<tr>
<td>( C_{m} )</td>
<td>( 1/L_{ox} )</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>( C_{gas} )</td>
<td>( C_{gas}WL )</td>
<td>1/S</td>
<td>1/S</td>
<td>1/S</td>
</tr>
<tr>
<td>( b_m, b_p )</td>
<td>( C_{ox}V/L )</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>( I_{av} )</td>
<td>( b_{n, p} V^2 )</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>Current Density</td>
<td>( I_{av}/Area )</td>
<td>S</td>
<td>S^3/U</td>
<td>S^3</td>
</tr>
<tr>
<td>( I_{on} )</td>
<td>( V_{DDP} )</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>( \text{Injection Depth} )</td>
<td>( W/C_{gas} )</td>
<td>1/S</td>
<td>1/S</td>
<td>1/S</td>
</tr>
<tr>
<td>( P' )</td>
<td>( I_{av}V )</td>
<td>1/S^2</td>
<td>1/S U^2</td>
<td>1</td>
</tr>
<tr>
<td>( \text{Power Density} )</td>
<td>( P/\text{Area} )</td>
<td>1</td>
<td>S^3 U^2</td>
<td>S^3</td>
</tr>
</tbody>
</table>
µProcessor Scaling

2X Growth in 1.96 Years!

P.Gelsinger: µProcessors for the New Millenium, ISSCC 2001

µProcessor Power

P.Gelsinger: µProcessors for the New Millenium, ISSCC 2001
μProcessor Performance

- Pentium® 4 proc
- Pentium® II & III proc
- Pentium® proc

2010 Outlook

- Performance 2X/16 months
  - 1 TIP (terra instructions/s)
  - 30 GHz clock
- Size
  - No of transistors: 2 Billion
  - Die: 40*40 mm
- Power
  - 10kW!!
  - Leakage: 1/3 active Power
Some interesting questions

- What will cause this model to break?
- When will it break?
- Will the model gradually slow down?
  » Power and power density
  » Leakage
  » Process Variation