Today’s lecture

- Power consumption (wrap-up)
- Inverter chain sizing
Announcements

- Homework 3 due today.
- Homework 4 will be posted later today.

Power Dissipation
Where Does Power Go in CMOS?

- Switching power
  - Charging capacitors
- Leakage power
  - Transistors are imperfect switches
- Short-circuit power
  - Both pull-up and pull-down on during transition
- Static currents
  - Biasing currents, in e.g. memory

Transistor Leakage

- Transistors that are supposed to be off - leak

Input at $V_{DD}$

Input at 0
The Sub-Micron MOS Transistor

- Threshold Variations
- Subthreshold Conduction
- Parasitic Resistances

Threshold Variations

Long-channel threshold

Threshold as a function of the length (for low $V_{DS}$)

Drain-induced barrier lowering (for low $L$)
**Sub-Threshold Conduction**

The Slope Factor

\[ I_D \sim I_0 e^{\frac{qV_{GS}}{n kT}}, \quad n = 1 + \frac{C_D}{C_{ox}} \]

S is \( \Delta V_{GS} \) for \( I_{D2}/I_{D1} = 10 \)

\[ S = n \left( \frac{kT}{q} \right) \ln(10) \]

Typical values for S: 60..100 mV/decade

---

**Inverter Delay**

- Minimum length devices, \( L = 0.25 \mu m \)
- Assume that for \( W_P = 2W_N = 2W \)
  - same pull-up and pull-down currents
  - approx. equal resistances \( R_N = R_P \)
  - approx. equal rise \( t_{pLH} \) and fall \( t_{pHL} \) delays
- Analyze as an RC network

\[
R_P = R_{unit} \left( \frac{W_P}{W_{unit}} \right)^{-1} \approx R_{unit} \left( \frac{W_N}{W_{unit}} \right)^{-1} = R_N = R_W
\]

Delay (D):

\[ t_{pHL} = (\ln 2) R_N C_L \quad t_{pLH} = (\ln 2) R_P C_L \]

Load for the next stage:

\[ C_{gin} = 3 \frac{W}{W_{unit}} C_{unit} \]
**Sub-Threshold** $I_D$ vs $V_{GS}$

$$I_D = I_0 e^{\frac{qV_{GS}}{nkT}} \left( 1 - e^{-\frac{qV_{DS}}{kT}} \right)$$

**Sub-Threshold** $I_D$ vs $V_{DS}$

$$I_D = I_0 e^{\frac{qV_{GS}}{nkT}} \left( 1 - e^{-\frac{qV_{DS}}{kT}} \right) \left( 1 + \lambda \cdot V_{DS} \right)$$
Sizing of an Inverter Chain

Subthreshold Leakage Component

- Leakage control is critical for low-voltage operation
Static Power Consumption

\[
P_{\text{stat}} = P_{(I_{\text{in}}=1)} \cdot V_{\text{dd}} \cdot I_{\text{stat}}
\]

Wasted energy …
Should be avoided in most cases, but could help reducing energy in others (e.g. sense amps)

Principles for Power Reduction

- Prime choice: Reduce voltage!
  - Recent years have seen an acceleration in supply voltage reduction
  - Design at very low voltages still open question (0.6 … 0.9 V by 2010!)
- Reduce switching activity
- Reduce physical capacitance
**Inverter Chain**

If \( C_L \) is given:
- How many stages are needed to minimize the delay?
- How to size the inverters?

May need some additional constraints.

**Inverter with Load**

\[ t_p = k R_w C_L \]

\( k \) is a constant, equal to 0.69
Assumptions: no load -> zero delay

\( W_{\text{unit}} = 1 \)
**Inverter with Load**

\[ C_p = 2C_{\text{unit}} \]

\[ C_N = C_{\text{unit}} \]

\[ 2W \]

\[ W \]

\[ C_{\text{int}} \]

\[ C_L \]

Delay formula:

\[ \text{Delay} = kR_W(C_{\text{int}} + C_L) = kR_W C_{\text{int}} + kR_W C_L = kR_W C_{\text{int}} (1 + \frac{C_L}{C_{\text{int}}}) = \text{Delay (Internal)} + \text{Delay (Load)} \]

**Delay Formula**

Delay \( \sim R_W (C_{\text{int}} + C_L) \)

\[ t_p = kR_W C_{\text{int}} (1 + \frac{C_L}{C_{\text{int}}}) = t_{p0} \left(1 + \frac{f}{\gamma}\right) \]

- \( C_{\text{int}} = \gamma C_{\text{gin}} \) with \( \gamma \approx 1 \)
- \( f = C_L/C_{\text{gin}} \) - effective fanout
- \( R = \frac{R_{\text{unit}}}{W} ; C_{\text{int}} = WC_{\text{unit}} \)
- \( t_{p0} = 0.69 R_{\text{unit}} C_{\text{unit}} \)
**Apply to Inverter Chain**

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

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

\[ t_p = \sum_{j=1}^{N} t_{p,j} = t_{p0} \sum_{i=1}^{N} \left( 1 + \frac{C_{\text{gin},j+1}}{\gamma C_{\text{gin},j}} \right), \quad C_{\text{gin},N+1} = C_L \]

---

**Optimal Tapering for Given \( N \)**

Delay equation has \( N - 1 \) unknowns, \( C_{\text{gin},2} - C_{\text{gin},N} \)

Minimize the delay, find \( N - 1 \) partial derivatives

Result: \( \frac{C_{\text{gin},j+1}}{C_{\text{gin},j}} = \frac{C_{\text{gin},j}}{C_{\text{gin},j-1}} \)

Size of each stage is the geometric mean of two neighbors

\[ C_{\text{gin},j} = \sqrt[2]{C_{\text{gin},j-1}C_{\text{gin},j+1}} \]

- each stage has the same effective fanout (\( C_{\text{out}}/C_{\text{in}} \))
- each stage has the same delay
**Optimum Delay and Number of Stages**

When each stage is sized by \( f \) and has same eff. fanout \( f \):

\[
f^N = F = C_L / C_{gin,1}
\]

Effective fanout of each stage:

\[
f = \sqrt[3]{N F}
\]

Minimum path delay

\[
t_p = N t_{p0} \left( 1 + \sqrt[3]{N F} / \gamma \right)
\]

**Example**

\[\begin{array}{c}
\text{In} \\
\downarrow C_1 \\
\end{array} \quad \begin{array}{c}
\text{1} \\
\downarrow f \\
\end{array} \quad \begin{array}{c}
f^2 \\
\text{Out} \\
\downarrow C_L = 8 C_1 \\
\end{array}\]

\( C_L / C_1 \) has to be evenly distributed across \( N = 3 \) stages:

\[
f = \sqrt[3]{8} = 2
\]
Optimum Number of Stages

For a given load, $C_L$ and given input capacitance $C_{in}$
Find optimal sizing $f$

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

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

$$\frac{\partial t_p}{\partial f} = \frac{t_{p0} \ln F}{\gamma} \cdot \ln f - \gamma f + \gamma f = 0$$

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

$$f = \exp(1 + \gamma/f)$$

---

Optimum Effective Fanout $f$

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

$$f = \exp(1 + \gamma/f)$$

$f_{opt} = 3.6$

for $\gamma = 1$
**Impact of Loading on $t_p$**

With Self-Loading $\gamma=1$

![Graph showing impact of loading on $t_p$]

**Normalized delay function of $F$**

\[ t_p = Nt_{p0} \left( 1 + \sqrt[3]{F / \gamma} \right) \]

<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>65</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)?

\[ C_{\text{tot}} = C_i + eC_i + \ldots + e^N C_i \]
\[ = C_i (1 + e + \ldots + e^N) \]
\[ = C_i + C_i e^N + C_i e (1 + e + \ldots + e^{N-2}) \]

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

e.g. \(C_L = 20\text{pF}; C_i = 50\text{fF} \rightarrow N=6\)
Fixed: 20pF
overhead: 11.66pF !!!
**Delay versus Area and Power**

![Graph showing delay versus area and power](image)

**Next Lecture**

- Technology Scaling
- Process Variations
- Wires