Lecture 10: The Inverter, A Dynamic View

Heads up

- This lecture
  - Inverter dynamic view
    - Reading assignment – Rabaey, et al, 5.4.1-5.4.3

- Next lecture
  - Designing fast logic
    - Reading assignment – Rabaey, et al, 6.2.1, 9.2.2, 9.2.3
Inverter Propagation Delay

- Propagation delay is proportional to the time-constant of the network formed by the pull-down (or pull-up) resistor and the load capacitance

\[ t_{pHL} = f(R_n, C_L) \]

\[ V_{DD} \]

\[ \begin{align*}
V_{out} &= 0 \\
R_n &\quad \quad C_L
\end{align*} \]

\[ t_{pHL} = \ln(2) R_{eqn} C_L = 0.69 R_{eqn} C_L \]

\[ t_{PLH} = \ln(2) R_{eqp} C_L = 0.69 R_{eqp} C_L \]

\[ V_{in} = V_{DD} \]

\[ t_p = \frac{(t_{pHL} + t_{PLH})}{2} = 0.69 \ C_L (R_{eqn} + R_{eqp})/2 \]

- To equalize rise and fall times make the on-resistance of the NMOS and PMOS approximately equal.
Inverter Transient Response

$V_{DD} = 2.5V$

$0.25\mu m$

$W/L_n = 1.5$

$W/L_p = 4.5$

$R_{eqn} = \frac{13 \, k\Omega}{1.5}$

$R_{eqp} = \frac{31 \, k\Omega}{4.5}$

$V_{in}$

$V_{out} (V)$

$t (sec) \times 10^{-10}$
Inverter Transient Response

$V_{DD} = 2.5V$

$0.25\mu\text{m}$

$W/L_n = 1.5$

$W/L_p = 4.5$

$R_{eqn} = 13\ k\Omega \,(\div 1.5)$

$R_{eqp} = 31\ k\Omega \,(\div 4.5)$

$t_{pHL} = 36\ \text{psec}$

$t_{pLH} = 29\ \text{psec}$

so

$t_p = 32.5\ \text{psec}$
Inverter Propagation Delay, Revisited

To see how a designer can optimize the delay of a gate have to expand the $R_{eq}$ in the delay equation

$$t_{pHL} = 0.69 \ R_{eq} \ C_L$$

$$= 0.69 \ (3/4 \ (C_L \ V_{DD})/I_{DSATn})$$

$$\approx 0.52 \ C_L / (W/L_n \ k'_n \ V_{DSATn})$$
Design for Performance

- Increase W/L ratio of the transistor
  - the most powerful and effective performance optimization tool in the hands of the designer
  - watch out for self-loading! – when the intrinsic capacitance dominates the extrinsic capacitance

- Reduce $C_L$
  - keep drain diffusions small
  - limit interconnect capacitance
  - limit fan-out

- Increase $V_{DD}$
  - trade-off energy for performance
  - increasing $V_{DD}$ above a certain level yields minimal improvements
  - reliability concerns enforce a firm upper bound on $V_{DD}$

![Graph showing normalized $t_p$ vs. $V_{DD}$]
Impacts of NMOS/PMOS Ratio

- So far have sized the PMOS and NMOS so that the $R_{eq}$’s match (ratio ~ 3)
  - symmetrical VTC
  - equal high-to-low and low-to-high propagation delays

- If speed is the only concern, reduce the width of the PMOS device!
  - widening the PMOS degrades the $t_{pHL}$ due to larger intrinsic capacitance

\[
\beta = \frac{W/L_p}{W/L_n}
\]

\[
r = \frac{R_{eqp}}{R_{eqn}} \text{ (resistance ratio of identically-sized PMOS and NMOS)}
\]

\[
\beta_{opt} = \sqrt[n]{r} \text{ when wiring capacitance is negligible}
\]
PMOS/NMOS Ratio Effects

\[ \beta = \frac{(W/L_p)}{(W/L_n)} \]

- \( \beta \) of 2.4 (= 31 \( k\Omega \)/13 \( k\Omega \)) gives symmetrical response
- \( \beta \) of 1.6 to 1.9 gives optimal performance
Device Sizing for Performance

- Divide capacitive load, $C_L$, into
  - $C_{\text{int}}$: intrinsic - diffusion and Miller effect ($C_g$)
  - $C_{\text{ext}}$: extrinsic - wiring and fanout

$$t_p = 0.69 \ R_{eq} \ C_{\text{int}} \ (1 + C_{\text{ext}}/C_{\text{int}}) = t_{p0} \ (1 + C_{\text{ext}}/C_{\text{int}})$$

- where $t_{p0} = 0.69 \ R_{eq} \ C_{\text{int}}$ is the intrinsic (unloaded) delay of the gate
Device Sizing for Performance

- Divide capacitive load, $C_L$, into
  - $C_{int}$: intrinsic - diffusion and Miller effect ($C_g$)
  - $C_{ext}$: extrinsic - wiring and fanout

$$t_p = 0.69 \frac{R_{eq} C_{int}}{1 + \frac{C_{ext}}{C_{int}}} = t_{p0} \left(1 + \frac{C_{ext}}{C_{int}}\right)$$

- where $t_{p0} = 0.69 \frac{R_{eq} C_{int}}{1 + \frac{C_{ext}}{C_{int}}}$ is the intrinsic (unloaded) delay of the gate

- Widening both PMOS and NMOS by a factor $S$
  - $R_{eq1} = \frac{R_{eq}}{S}$, $C_{int1} = S C_{int}$

$$t_p = \frac{R_{eq} C_{int}}{1 + \frac{C_{ext}}{C_{int}}}$$

- $t_{p0}$ is independent of the sizing of the gate; with no load the drive of the gate is totally offset by the increased capacitance

- any $S$ sufficiently larger than $\frac{C_{ext}}{C_{int}}$ yields the best performance gains with least area impact
The majority of the improvement is already obtained for $S = 5$. Sizing factors larger than 10 barely yield any extra gain (and cost significantly more area).
**Impact of Fanout on Delay**

\[ t_p = t_{p0} \left( 1 + \frac{C_{\text{ext}}}{C_{\text{int}}} \right) \]

- Extrinsic capacitance, \( C_{\text{ext}} \), is a function of the fanout of the gate - the larger the fanout, the larger the external load.

- First determine the **input loading** effect of the inverter. Both \( C_g \) and \( C_{\text{int}} \) are proportional to the gate sizing, so \( C_{\text{int}} = \gamma C_g \), \( \gamma \) is independent of gate sizing and

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

- The delay of an inverter is a function of the ratio between its external load capacitance and its input gate capacitance, or the **gate's effective fan-out** \( f \)

  \[ f = \frac{C_{\text{ext}}}{C_g} \]
Inverter Chain

- Real goal is to minimize the delay through an inverter chain

![Inverter Chain Diagram]

- The delay of the j-th inverter stage is
  \[ t_{p,j} = t_{p0} \left( 1 + \frac{C_{g,j+1}}{\gamma C_{g,j}} \right) = t_{p0} \left( 1 + \frac{f_j}{\gamma} \right) \]

- And
  \[ t_{p} = t_{p1} + t_{p2} + \ldots + t_{pN} \]

- So
  \[ t_{p} = \sum t_{p,j} = t_{p0} \sum \left( 1 + \frac{C_{g,j+1}}{\gamma C_{g,j}} \right) \]

- If \( C_L \) is given
  - How should the inverters be sized?
  - How many stages are needed to minimize the delay?
Sizing the Inverters in the Chain of N inverters

- The optimum size of each inverter is the geometric mean of its neighbors – meaning that if each inverter is sized up by the same factor $f$ wrt the preceding gate, it will have the same effective fan-out and the same delay

$$f = \sqrt{\frac{C_L}{C_{g,1}}} = \sqrt{F}$$

where the overall effective fan-out of the circuit is

$$F = \frac{C_L}{C_{g,1}}$$

and the minimum delay through the inverter chain is

$$t_p = N \, t_{p0} \left( 1 + \left( \sqrt{N} \frac{\sqrt{F}}{\gamma} \right) \right)$$

- The relationship between $t_p$ and $F$ is linear for one inverter, square root for two, etc.
Example of Inverter Chain Sizing

\[ \frac{C_L}{C_{g,1}} \text{ has to be evenly distributed over } N = 3 \text{ inverters} \]

\[ F = \frac{C_L}{C_{g,1}} = \frac{8}{1} \]

\[ f = \]
Example of Inverter Chain Sizing

- $C_L/C_{g,1}$ has to be evenly distributed over $N = 3$ inverters.

$$F = \frac{C_L}{C_{g,1}} = \frac{8}{1}$$

$$f = \sqrt[3]{8} = 2$$
Determining N: Optimal Number of Inverters

- What is the optimal value for N given F (= \( f^N \))?
  - if the number of stages is too large, the intrinsic delay of the stages becomes dominate
  - if the number of stages is too small, the effective fan-out of each stage becomes dominate

The optimum N is found by differentiating the minimum delay expression divided by the number of stages and setting the result to 0, giving

\[ \gamma + \sqrt[N]{F} - \left( \frac{\sqrt[N]{F \ln(F)}}{N} \right) = 0 \quad \text{and} \quad f = e^{(1 + \gamma/f)} \]

- For \( \gamma = 0 \) (ignoring self-loading) \( N = \ln(F) \)
  and the effective-fan out (tapering factor) is \( f = e = 2.718 \)

- For \( \gamma = 1 \) (the typical case) \( N = \ln(F) - 1 \)
  and the effective fan-out (tapering factor) is \( f = 3.6 \sim 4 \)
Choosing $f$ larger than optimum has little effect on delay and reduces the number of stages (and area).

- So it is common practice to use $f = 4$ (for $\gamma = 1$) and reduce $N$
- Too many stages has a substantial negative impact on delay
Example of Inverter (Buffer) Staging

\[ F = 64 \quad N = \ln(F) - 1 = 3.16 \]

\[ C_{g,1} = 1 \quad C_L = 64 \ C_{g,1} \]

\[ \begin{array}{c}
1 \quad 8 \\
C_{g,1} = 1 \\
\end{array} \quad \begin{array}{c}
2 \quad 8 \quad 18 \\
\end{array} \]

\[ \begin{array}{c}
1 \quad 4 \quad 16 \\
C_{g,1} = 1 \\
\end{array} \quad \begin{array}{c}
3 \quad 4 \quad 15 \\
\end{array} \]

\[ \begin{array}{c}
1 \quad 2.8 \quad 8 \quad 22.6 \\
C_{g,1} = 1 \\
\end{array} \quad \begin{array}{c}
4 \quad 2.8 \quad 15.3 \\
\end{array} \]
Impact of Buffer Staging for Large $C_i$

<table>
<thead>
<tr>
<th>$F$ ($\gamma = 1$)</th>
<th>Unbuffered</th>
<th>Two Stage Chain</th>
<th>Opt. 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>1,000</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>

- Impressive speed-ups with optimized cascaded inverter chain for very large capacitive loads.
Input Signal Rise/Fall Time

- In reality, the input signal changes gradually (and both PMOS and NMOS conduct for a brief time). This affects the current available for charging/discharging $C_L$ and impacts propagation delay.

- $t_p$ increases linearly with increasing input slope, $t_s$, once $t_s > t_p$.

- $t_s$ is due to the limited driving capability of the preceding gate for a minimum-size inverter with a fan-out of a single gate.
A gate is never designed in isolation: its performance is affected by both the fan-out and the driving strength of the gate(s) feeding its inputs.

\[ t_p^i = t_{\text{step}}^i + \eta \cdot t_{\text{step}}^{i-1} \quad (\eta \approx 0.25) \]

- Keep signal rise times smaller than or equal to the gate propagation delays.
  - good for performance
  - good for power consumption

- Keeping rise and fall times of the signals small and of approximately equal values is one of the major challenges in high-performance designs - slope engineering.
Delay with Long Interconnects

- When gates are farther apart, wire capacitance and resistance can no longer be ignored.

\[
V_{in} \quad (r_w, c_w, L) \quad V_{out}
\]

\[
t_p = 0.69R_{dr}C_{int} + (0.69R_{dr} + 0.38R_w)C_w + 0.69(R_{dr} + R_w)C_{fan}
\]

where \( R_{dr} = (R_{eqn} + R_{eqp})/2 \)

\[
= 0.69R_{dr}(C_{int} + C_{fan}) + 0.69(R_{dr}c_w + r_wC_{fan})L + 0.38r_wc_wL^2
\]

- Wire delay rapidly becomes the dominate factor (due to the quadratic term) in the delay budget for longer wires.
Next Lecture and Reminders

- Next lecture
  - Designing fast logic
    - Reading assignment – Rabaey, et al, 6.2.1,9.2.2,9.2.3