The Inverter

Revised from Digital Integrated Circuits, © Jan M. Rabaey et al, 2003
Propagation Delay
CMOS Inverter Propagation Delay

\[
t_{pHL} = f(R_{on} \cdot C_L) = 0.69 \ R_{on} C_L
\]

\[
\text{ln}(0.5) = 0.36
\]

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

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

\[
R_{on}\ C_L
\]

\[
t
\]

\[
V_{out}
\]

\[
V_{DD}
\]

\[
0.5
\]

\[
0.36
\]
MOS transistor model for simulation
Computing the Capacitances

Consider each capacitor individually is almost impossible for manual analysis. What capacitors count in $C_L$?
Computing the Capacitances

- NMOS and PMOS transistor are either in cutoff or saturation mode during at least the first half (50%) of the output transient.

- So, the only contributions to $C_{gd}$ are the overlap capacitance, since channel capacitance occurs between either Gate-Body for transistors in cutoff region or Gate-Source for transistors in saturation region.
The Miller Effect

The lumped capacitor model requires the floating $C_{gd1}$ capacitor be replaced by a capacitor to GND using Miller effect.

“A capacitor experiencing identical but opposite voltage swings at both its terminals can be replaced by a capacitor to ground, whose value is two times the original value.”
The Miller Effect

Consider the situation that an impedance is connected between input and output of an amplifier.

- The same current flows from (out) the top input terminal if an impedance $Z_{in, Miller}$ is connected across the input terminals.
- The same current flows to (in) the top output terminal if an impedance $Z_{out, Miller}$ is connected across the output terminal.
- This is known as Miller Effect.
- Two important notes to apply Miller Effect:
  - There should be a common terminal for input and output.
  - The gain in the Miller Effect is the gain after connecting feedback impedance $Z_f$. 

\[
I_f = \frac{V_i(1-A_v)}{Z_f} + V_f - \quad \text{(a)}
\]

\[
V_i = A_vV_o + \quad \text{Amplifier}
\]

\[
Z_{in, Miller} = \frac{Z_f}{1-A_v} \quad \text{(b)}
\]

\[
Z_{out, Miller} = \frac{Z_fA_v}{A_v-1}
\]

Graphs from Prentice Hall
Computing the Capacitances

<table>
<thead>
<tr>
<th>Capacitor</th>
<th>Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_{gd1}$</td>
<td>$2 \times CGD0 \times W_n$</td>
</tr>
<tr>
<td>$C_{gd2}$</td>
<td>$2 \times CGD0 \times W_p$</td>
</tr>
<tr>
<td>$C_{db1}$</td>
<td>$K_{eqn} \times (AD_n \times CJ + PD_n \times CJSW)$</td>
</tr>
<tr>
<td>$C_{db2}$</td>
<td>$K_{eqp} \times (AD_p \times CJ + PD_p \times CJSW)$</td>
</tr>
<tr>
<td>$C_{g3}$</td>
<td>$C_{ox} \times W_n \times L_n$</td>
</tr>
<tr>
<td>$C_{g4}$</td>
<td>$C_{ox} \times W_p \times L_p$</td>
</tr>
<tr>
<td>$C_w$</td>
<td>From Extraction</td>
</tr>
<tr>
<td>$C_L$</td>
<td>$\Sigma$</td>
</tr>
</tbody>
</table>

The capacitance between drain and bulk, $C_{db1}$ and $C_{db2}$, are due to the reverse-biased pn-junction.

Such a capacitor is, unfortunately, quite nonlinear and depends heavily on the applied voltage.

In Chapter 3 we replaced the nonlinear capacitor by a linear one with the same change in charge for the voltage range of interest. A multiplication factor, $K_{eq}$, is introduced to relate the linearized capacitor to the value of the junction capacitance under zero-bias conditions (usually in the range of 0.6 to 0.9).
Junction Capacitance

\[ C_j = \frac{C_{j0}}{(1 - V_D / \phi_0)^m} \]

- \( m = 0.5 \): abrupt junction
- \( m = 0.33 \): linear junction
Linearizing the Junction Capacitance

Replace non-linear capacitance by large-signal equivalent linear capacitance which displaces equal charge over voltage swing of interest

\[ C_{eq} = \frac{\Delta Q_j}{\Delta V_D} = \frac{Q_j(V_{\text{high}}) - Q_j(V_{\text{low}})}{V_{\text{high}} - V_{\text{low}}} = K_{eq} C_{j0} \]

\[ K_{eq} = \frac{-\phi_0^m}{(V_{\text{high}} - V_{\text{low}})(1 - m)}\left[(\phi_0 - V_{\text{high}})^{1-m} - (\phi_0 - V_{\text{low}})^{1-m}\right] \]
Table 6: Contact to Active

<table>
<thead>
<tr>
<th>Rule</th>
<th>Description</th>
<th>Lambda</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1</td>
<td>Exact contact size</td>
<td>2 x 2</td>
</tr>
<tr>
<td>6.2</td>
<td>Minimum active overlap</td>
<td>3</td>
</tr>
<tr>
<td>6.3</td>
<td>Minimum contact spacing</td>
<td>3</td>
</tr>
<tr>
<td>6.4</td>
<td>Minimum spacing to gate of transistor</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 7: Metal1

<table>
<thead>
<tr>
<th>Rule</th>
<th>Description</th>
<th>Lambda</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.1</td>
<td>Minimum width</td>
<td>3</td>
</tr>
<tr>
<td>7.2</td>
<td>Minimum spacing</td>
<td>3</td>
</tr>
<tr>
<td>7.3</td>
<td>Minimum overlap of any contact</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3: Poly

<table>
<thead>
<tr>
<th>Rule</th>
<th>Description</th>
<th>Lambda</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Minimum width</td>
<td>2</td>
</tr>
<tr>
<td>3.2</td>
<td>Minimum spacing</td>
<td>3</td>
</tr>
<tr>
<td>3.3</td>
<td>Minimum spacing over active</td>
<td>3</td>
</tr>
<tr>
<td>3.4</td>
<td>Minimum gate extension of active</td>
<td>2</td>
</tr>
<tr>
<td>3.5</td>
<td>Minimum active extension of poly</td>
<td>3</td>
</tr>
<tr>
<td>3.6</td>
<td>Minimum field polyc to active</td>
<td>1</td>
</tr>
</tbody>
</table>
**Junction Capacitance $L_s$ (from Ch. 3)**

Channel-stop implant

$N_A$

Side wall

Source $N_D$

Bottom

Channel

Substrate $N_A$

W

$L_S$

$x_j$

Junction capacitance per unit length

$C_{\text{diff}} = C_{\text{bottom}} + C_{\text{sw}} = C_j \times \text{AREA} + C_{j_{\text{sw}}} \times \text{PERIMETER}$

$= C_j L_S W + C_{j_{\text{sw}}} (2L_S + W)$

Junction capacitance per unit area

No channel side

Inverter

© Digital Integrated Circuits

2nd
Computation of all capacitors

<table>
<thead>
<tr>
<th>Capacitor</th>
<th>Expression</th>
<th>Value (fF) (H→L)</th>
<th>Value (fF) (L→H)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_{gd1}$</td>
<td>$2 \cdot CGD_{0_n} \cdot W_n$</td>
<td>0.23</td>
<td>0.23</td>
</tr>
<tr>
<td>$C_{gd2}$</td>
<td>$2 \cdot CGD_{0_p} \cdot W_p$</td>
<td>0.61</td>
<td>0.61</td>
</tr>
<tr>
<td>$C_{db1}$</td>
<td>$K_{eqn} \cdot A_{D_n} \cdot CJ + K_{eqswn} \cdot P_{D_n} \cdot CJSW$</td>
<td>0.66</td>
<td>0.90</td>
</tr>
<tr>
<td>$C_{db2}$</td>
<td>$K_{eqp} \cdot A_{D_p} \cdot CJ + K_{eqswp} \cdot P_{D_p} \cdot CJSW$</td>
<td>1.5</td>
<td>1.15</td>
</tr>
<tr>
<td>$C_g3$</td>
<td>$(CGD_{0_n} + CGSO_{0_n}) \cdot W_n + C_{ox} \cdot W_n \cdot L_n$</td>
<td>0.76</td>
<td>0.76</td>
</tr>
<tr>
<td>$C_g4$</td>
<td>$(CGD_{0_p} + CGSO_{0_p}) \cdot W_p + C_{ox} \cdot W_p \cdot L_p$</td>
<td>2.28</td>
<td>2.28</td>
</tr>
<tr>
<td>$C_w$</td>
<td>From Extraction</td>
<td>0.12</td>
<td>0.12</td>
</tr>
<tr>
<td>$C_L$</td>
<td>$\sum$</td>
<td>6.1</td>
<td>6.0</td>
</tr>
</tbody>
</table>

$C_{db}$ will be **slightly** different in $L$-to-$H$ and $H$-to-$L$, why? (the pn junction reverse bias voltage range)
**Transient Response**

*C<sub>gd</sub> directly couples the steep input change before the circuit can even start to react to the changes at input (potential forward bias the pn junction)*

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

\[
R_{eq} = \frac{1}{V_{DD}/2} \int_{V_{DD}/2}^{V_{DD}} \frac{V}{I_{DSAT}(1 + \lambda V)} dV \approx \frac{3}{4} \frac{V_{DD}}{I_{DSAT}} \left(1 - \frac{7}{9} \lambda V_{DD}\right)
\]

with \[ I_{DSAT} = \frac{k}{L} \left(V_{DD} - V_T\right) V_{DSAT} - \frac{V_{DSAT}^2}{2} \]
Low-to-High and High-to-Low delay

- It is desired to have identical propagation delays for both rising and falling inputs.

- Equal delay requires equal equivalent on-resistance, thus equal current $I_{DAST}$ (neglecting the channel length modulation)

- This demands almost the same requirements for a $V_m$ at $V_{DD}/2$. Why?
Requirements for equal delay

\[ I_{DSAT} = k \cdot \frac{W}{L} \left( (V_{DD} - V_T)V_{DSAT} - \frac{V_{DSAT}^2}{2} \right) \]

\[ I_{Dn} = k_n \cdot \left( \frac{W}{L} \right)_n V_{DSATn} \left( (V_{DD} - V_{Tn}) - \frac{V_{DSATn}}{2} \right) \]

\[ I_{Dp} = k_p \cdot \left( \frac{W}{L} \right)_p V_{DSATp} \left( (V_{DD} - V_{Tp}) - \frac{V_{DSATp}}{2} \right) \]

Assume \( V_{DD} >> V_{Tp} + V_{DSATp} / 2 \)

then \[ \frac{k_p V_{DSATp} (W / L)_p}{k_n V_{DSATn} (W / L)_n} = \frac{k_p V_{DSATp}}{k_n V_{DSATn}} = 1 \]

This is exactly the formerly defined parameter \( r \) (last lecture)
Design for delay performance

- Keep capacitances small
  - careful layout, e.g. to keep drain diffusion as small as possible

- Increase transistor sizes
  - watch out for self-loading! When intrinsic capacitance starts to dominate the extrinsic ones

- Increase VDD (????)
Delay as a function of $V_{DD}$

\[ t_{PHL} = 0.69 \frac{3C_L V_{DD}}{4I_{DSATn}} = 0.52 \frac{C_L V_{DD}}{(W/L)_{n} k'_n V_{DSATn}(V_{DD} - V_{Tn} - V_{DSATn}/2)} \]

For fixed (W/L)

Recall a range of low voltage is able to give even better voltage transfer characteristic.
Device Sizing

(for fixed load and \(V_{DD}\))

Self-loading effect: Intrinsic capacitances dominate
Widening PMOS improves the L-H delay by increasing the charge current, but it also degrades the H-L by giving a larger parasitic capacitance. Considering average is more meaningful!!
Delay Definitions

![Diagram showing delay definitions and timing parameters for an inverter circuit.](image)
Impact of Rise Time on Delay

\[ t_{pHL} = \sqrt{t_{pHL}^2 + \left(\frac{t_{rise}}{2}\right)^2} \]
Custom design process: An inverter design example
1. Schematic design
2. Layout design
Step 1: define Nwell (for PMOS)
Step 2: define pselect (for PMOS location)
Step 3: define active region (for PMOS)
Step 4: define poly (gate for PMOS)
Step 5: define contacts (for PMOS)
Step 6: define Vdd and connect source (of PMOS) to it
Step 7: make at least one Nwell contact
Step 8: create NMOS (repeat similar steps before except you do not need make Nwell)
step9: make input and output connections
2. DRC (Design Rule Check)

In the image, there is a screenshot of a program output showing DRC results. The program is ICfb, and the log file is /home/research/thua/CDS.log. The log contains several lines of text, including commands and output from the DRC process. The summary of rule violations for the cell "inverter layouttest" is shown, indicating 9 matching markers in view: "inverter".

The summary includes:
- Violated Rules:
  1. (SCMOS Rule 5.2.b) poly enclosure of contact: 0.15 um
  8. (SCMOS Rule 7.3) metall enclosure of contact: 0.15 um

The output also includes instructions for using the software, such as:
- Mouse L: Enter Point
- M: Pop-up Menu
- R: Toggle L90 X/Y

The page is part of a Digital Integrated Circuits 2nd edition, with the section title "Inverter" at the bottom right corner.
Correct error if there is any
After correction
3. LVS (Layout versus Schematic)

The net-lists match.
4. Extract Layout parasitics and post-layout simulation
CMOS Inverter

The CMOS inverter is a fundamental building block in digital integrated circuits. It consists of a PMOS and a NMOS transistor. In this circuit, the PMOS transistor is driven by the input signal, while the NMOS transistor is connected to the ground. The output voltage is controlled by the voltage across the two transistors.

The circuit diagram shows the basic topology of a CMOS inverter, with input (In) and output (Out) terminals. The voltage sources are labeled as $V_{DD}$ for the power supply and GND for the ground. The transistors are labeled PMOS and NMOS, with the PMOS transistor shaded in green and the NMOS transistor shaded in red.

The physical layout of the CMOS inverter is also depicted, showing the metal layers and contacts. The contacts are highlighted in red, and the PMOS and NMOS transistors are shown in green and red, respectively.

This CMOS inverter circuit is a key component in many digital circuits, enabling the implementation of logic operations such as AND, OR, and NOT.
Two Inverters

Layout preference:
*Share power and ground*
*Abut cells*

Connect in Metal

![Diagram of two inverters with metal connection](image)
Impact of Process Variations (DFM)

![Graph showing the impact of process variations on inverter performance]

- Good PMOS
- Bad NMOS
- Good NMOS
- Bad PMOS

Nominal

Inverter
Inverter Sizing for delay
Inverter with Load

$W$ means the size is increased by a factor of $W$ with respect to the minimum size.

![Inverter Diagram]

$$\text{Delay} = kR_W(C_{\text{int}} + C_L) = kR_W C_{\text{int}} + kR_W C_L$$

$$= \text{Delay (Internal)} + \text{Delay (Load)}$$

$$= kR_W C_{\text{int}}(1 + C_L / C_{\text{int}})$$
Delay as function of size

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

\[
R_W = \frac{R_{unit}}{W} ; C_{int} = W C_{unit}
\]

\[
t_p = t_{p0}(1 + \frac{C_L}{(WC_{unit})})
\]

\[
t_{p0} = 0.69R_{unit}C_{unit}
\]

- **Intrinsic delay is fixed and independent of size** \(W\)

- Making \(W\) large yields better performance gain, eliminating the impact of external load and reducing the delay to intrinsic only. But smaller gain at penalty of silicon area if \(W\) is too large!
Delay Formula

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

$$t_p = k R_W C_{\text{int}} \left( 1 + C_L / C_{\text{int}} \right) = t_{p0} \left( 1 + f / \gamma \right)$$

$$C_{\text{int}} = \gamma C_{gin} \text{ with } \gamma \approx 1 \text{ for modern technology}$$

(see page 199 in book or Slid 14 for an example)

$$C_{gin} : \text{input gate capacitance}$$

$$C_L = f C_{gin} \text{ - effective fanout}$$

This formula maps the intrinsic capacitor and load capacitor as functions of a **common capacitor**, which is the gate capacitance of the minimum-size inverter.
Single inverter versus inverter chain

- Gate sizing for an isolated gate is not really meaningful. Realistic chips always have a long chain of gates.

- So, a more relevant and realistic problem is to determine the optimal sizing for a chain of gates.
Inverter Chain

If $C_L$ is given:

- How many stages are needed to minimize the delay?

- How to size the inverters?
Apply to Inverter Chain (fixed N stages)

Unit size (minimum size) inverter

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

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

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

- Delay equation has $N - 1$ unknowns, $C_{gin,2} - C_{gin,N}$

- Minimize the delay, find $N - 1$ partial derivatives equated to 0

- Result: $C_{gin,j+1}/C_{gin,j} = C_{gin,j}/C_{gin,j-1}$

- Size of each stage is the geometric mean of two neighbors

$$C_{gin,j} = \sqrt{C_{gin,j-1}C_{gin,j+1}}$$

- each stage has the same effective fanout
- each stage has the same delay
**Optimum Delay for fixed N stages**

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

$$f^N = F = \frac{C_L}{C_{gin,1}} \quad \text{effective fanout of the overall circuit}$$

Effective fanout of each stage:

$$f = \sqrt[2N]{F}$$

Minimum path delay

$$t_p = \sum_{j=1}^{N} t_{p,j} = t_{p_0} \sum_{i=1}^{N} \left( 1 + \frac{C_{gin,j+1}}{\gamma C_{gin,j}} \right) = Nt_{p_0} \left( 1 + \sqrt[2N]{F} / \gamma \right)$$
Example

$C_L = 8C_1$

$C_L/C_1$ 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 $f$

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

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

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

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

$$f = \exp\left(1 + \gamma / f\right)$$
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 $\gamma = 1$
Impact of introducing buffers

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

\[ f_{opt} = 4 \]

<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>
Itanium has 6 integer execution units like this.