CMPEN 411
VLSI Digital Circuits
Spring 2009

Lecture 09: Resistance & Inverter Dynamic View

CMOS Inverter: Dynamic

- Transient, or dynamic, response determines the maximum speed at which a device can be operated.

\[ V_{DD} \]

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

\[ V_{out} = 0 \]

\[ R_n \]

\[ C_L \]

\[ t_{pHL} = f(R_n, C_L) \]
Review: Sources of Capacitance

intrinsic MOS transistor capacitances
extrinsic MOS transistor (fanout) capacitances
wiring (interconnect) capacitance
Review: Components of $C_L$ (0.25 $\mu$m)

<table>
<thead>
<tr>
<th>C Term</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>$2 \ C_{on} \ W_n$</td>
<td>0.23</td>
<td>0.23</td>
</tr>
<tr>
<td>$C_{GD2}$</td>
<td>$2 \ C_{op} \ W_p$</td>
<td>0.61</td>
<td>0.61</td>
</tr>
<tr>
<td>$C_{DB1}$</td>
<td>$K_{eqbpn} \ AD_n \ C_j + K_{eqswn} \ PD_n \ C_{jsw}$</td>
<td>0.66</td>
<td>0.90</td>
</tr>
<tr>
<td>$C_{DB2}$</td>
<td>$K_{eqbpp} \ AD_p \ C_j + K_{eqswp} \ PD_p \ C_{jsw}$</td>
<td>1.5</td>
<td>1.15</td>
</tr>
<tr>
<td>$C_{G3}$</td>
<td>$(2 \ C_{on}) \ W_n + C_{ox} \ W_n L_n$</td>
<td>0.76</td>
<td>0.76</td>
</tr>
<tr>
<td>$C_{G4}$</td>
<td>$(2 \ C_{op}) \ W_p + C_{ox} \ W_p L_p$</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>

Cint $\sim$=Cext
Sources of Resistance

MOS structure resistance - $R_{on}$
Source and drain resistance
Contact (via) resistance
Wiring resistance
The simplest model assumes the transistor is a switch with an infinite “off” resistance and a finite “on” resistance $R_{on}$.

\[ V_{GS} \geq V_T \]

\[ \text{S} \quad \Theta \quad \bigg\| \quad \text{D} \quad \bigg\| R_{on} \]

However $R_{on}$ is nonlinear, so use instead the average value of the resistances, $R_{eq}$, at the end-points of the transition ($V_{DD}$ and $V_{DD}/2$).

\[ R_{eq} = \frac{1}{2} (R_{on}(t_1) + R_{on}(t_2)) \]

\[ R_{eq} = \frac{3}{4} V_{DD}/I_{DSAT} (1 - 5/6 \lambda V_{DD}) \]
Equivalent MOS Structure Resistance

- The on resistance is inversely proportional to W/L. Doubling W halves $R_{eq}$.

- For $V_{DD} \gg V_T + V_{DSAT}/2$, $R_{eq}$ is independent of $V_{DD}$ (see plot). Only a minor improvement in $R_{eq}$ occurs when $V_{DD}$ is increased (due to channel length modulation).

- Once the supply voltage approaches $V_T$, $R_{eq}$ increases dramatically.

<table>
<thead>
<tr>
<th>$V_{DD}$ (V)</th>
<th>1</th>
<th>1.5</th>
<th>2</th>
<th>2.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMOS (kΩ)</td>
<td>35</td>
<td>19</td>
<td>15</td>
<td>13</td>
</tr>
<tr>
<td>PMOS (kΩ)</td>
<td>115</td>
<td>55</td>
<td>38</td>
<td>31</td>
</tr>
</tbody>
</table>

$R_{eq}$ (for W/L = 1), for larger devices divide $R_{eq}$ by W/L.
Source and Drain Resistance

\[ R_{S,D} = \left( \frac{L_{S,D}}{W} \right) R_{\square} \]

where \( L_{S,D} \) is the length of the source or drain diffusion

\( R_{\square} \) is the sheet resistance of the source or drain diffusion (20 to 100 Ohm/\( \square \))

- More pronounced with scaling since junctions are shallower

- With silicidation \( R_{\square} \) is reduced to the range 1 to 4 Ohm/\( \square \)
Contact Resistance

- Transitions between routing layers (contacts through via’s) add extra resistance to a wire
  - keep signals wires on a single layer whenever possible
  - avoid excess contacts
  - reduce contact resistance by making vias larger (beware of current crowding that puts a practical limit on the size of vias) or by using multiple minimum-size vias to make the contact

- Typical contact resistances, $R_C$, (minimum-size)
  - 5 to 20 Ω for metal or poly to n+, p+ diffusion and metal to poly
  - 1 to 5 Ω for metal to metal contacts

- More pronounced with scaling since contact openings are smaller
Wire Resistance

\[ R = \frac{\rho L}{A} = \frac{\rho L}{H W} \]

Sheet Resistance \( R_{\square} \)

\[ R_{1\square} = R_{2\square} \]

<table>
<thead>
<tr>
<th>Material</th>
<th>Sheet Res. ((\Omega/\square))</th>
</tr>
</thead>
<tbody>
<tr>
<td>n, p well diffusion</td>
<td>1000 to 1500</td>
</tr>
<tr>
<td>n+, p+ diffusion</td>
<td>50 to 150</td>
</tr>
<tr>
<td>n+, p+ diffusion with silicide</td>
<td>3 to 5</td>
</tr>
<tr>
<td>polysilicon</td>
<td>150 to 200</td>
</tr>
<tr>
<td>polysilicon with silicide</td>
<td>4 to 5</td>
</tr>
<tr>
<td>Aluminum</td>
<td>0.05 to 0.1</td>
</tr>
</tbody>
</table>

Material

<table>
<thead>
<tr>
<th>Material</th>
<th>(\rho(\Omega\text{-m}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Silver (Ag)</td>
<td>1.6 \times 10^{-8}</td>
</tr>
<tr>
<td>Copper (Cu)</td>
<td>1.7 \times 10^{-8}</td>
</tr>
<tr>
<td>Gold (Au)</td>
<td>2.2 \times 10^{-8}</td>
</tr>
<tr>
<td>Aluminum (Al)</td>
<td>2.7 \times 10^{-8}</td>
</tr>
<tr>
<td>Tungsten (W)</td>
<td>5.5 \times 10^{-8}</td>
</tr>
</tbody>
</table>
Skin Effect

- At high frequency, currents tend to flow primarily on the surface of a conductor with the current density falling off exponentially with depth into the wire.

\[ \delta = \sqrt{\frac{\rho}{\pi f \mu}} \]

where \( f \) is frequency
\( \mu = 4\pi \times 10^{-7} \text{ H/m} \)

so the overall cross section is \( \sim 2(W+H)\delta \)

- The onset of skin effect is at \( f_s \) - where the skin depth is equal to half the largest dimension of the wire.

\[ f_s = \frac{4 \rho}{\pi \mu (\max(W,H))^2} \]

- An issue for high frequency, wide (tall) wires (i.e., clocks!)
A 30% increase in resistance is observe for 20 μm Al wires at 1 GHz (versus only a 1% increase for 1 μm wires)
The Wire

transmitters  receivers

schematic  physical
Wire Models

- Interconnect parasitics (capacitance, resistance, and inductance)
  - reduce reliability
  - affect performance and power consumption

All-inclusive (C,R,I) model

Capacitance-only
Parasitic Simplifications

- Inductive effects can be ignored
  - if the resistance of the wire is substantial enough (as is the case for long Al wires with small cross section)
  - if the rise and fall times of the applied signals are slow enough

- When the wire is short, or the cross-section is large, or the interconnect material has low resistivity, a capacitance only model can be used

- When the separation between neighboring wires is large, or when the wires run together for only a short distance, interwire capacitance can be ignored and all the parasitic capacitance can be modeled as capacitance to ground
Simulated Wire Delays

\[ V_{in} \quad \int \quad V_{out} \]

\[ \frac{L}{10} \quad \frac{L}{4} \quad \frac{L}{2} \quad L \]

\[ \text{time (nsec)} \]

\[ \text{voltage (V)} \]

Sp09  CMPEN 411  L09  S.16
Wire Delay Models

- Ideal wire
  - same voltage is present at every segment of the wire at every point in time - at equi-potential
  - only holds for very short wires, i.e., interconnects between very nearest neighbor gates

- Lumped C model
  - when only a single parasitic component (C, R, or L) is dominant the different fractions are lumped into a single circuit element
    - When the resistive component is small and the switching frequency is low to medium, can consider only C; the wire itself does not introduce any delay; the only impact on performance comes from wire capacitance
  - good for short wires; pessimistic and inaccurate for long wires
Wire Delay Models, con’t

- **Lumped RC model**
  - total wire resistance is lumped into a single R and total capacitance into a single C
  - good for short wires; pessimistic and inaccurate for long wires

- **Distributed RC model**
  - circuit parasitics are distributed along the length, L, of the wire
    - c and r are the capacitance and resistance per unit length

  ![Diode Circuit Diagram]

  - Delay is determined using the *Elmore delay* equation
    \[ \tau_{Di} = \sum_{k=1}^{N} c_k r_{ik} \]
RC Tree Definitions

- **RC tree characteristics**
  - A unique resistive path exists between the source node and any node of the network
    - Single input (source) node, \( s \)
    - All capacitors are between a node and GND
    - No resistive loops
  - **Path resistance** (sum of the resistances on the path from the input node to node \( i \))
    \[
    r_{ii} = \sum_{j=1}^{i} r_j \Rightarrow (r_j \in \text{path}(s \rightarrow i))
    \]
  - **Shared path resistance** (resistance shared along the paths from the input node to nodes \( i \) and \( k \))
    \[
    r_{ik} = \sum_{j=1}^{N} r_j \Rightarrow (r_j \in \text{path}(s \rightarrow i) \cap \text{path}(s \rightarrow k))
    \]

- **A typical wire is a chain network** with (simplified) Elmore delay of
  \[
  \tau_{DN} = \sum_{i=1}^{N} c_i r_{ii}
  \]
Chain Network Elmore Delay

Elmore delay equation

\[ \tau_{DN} = \sum c_i r_{ii} = \sum c_i \sum r_j \]
Chain Network Elmore Delay

\[ \tau_{D1} = c_1 r_1 \quad \tau_{D2} = c_1 r_1 + c_2 (r_1 + r_2) \]

\[ \tau_{D_i} = c_1 r_1 + c_2 (r_1 + r_2) + \ldots + c_i (r_1 + r_2 + \ldots + r_i) \]

Elmore delay equation

\[ \tau_{DN} = \sum c_i r_{ii} = \sum c_i \sum r_j \]

\[ \tau_{D_i} = c_1 r_{eq} + 2c_2 r_{eq} + 3c_3 r_{eq} + \ldots + ic_i r_{eq} \]
Elmore Delay Models Uses

- Modeling the delay of a wire
- Modeling the delay of a series of pass transistors
- Modeling the delay of a pull-up and pull-down networks
Distributed RC Model for Simple Wires

- A length L RC wire can be modeled by N segments of length L/N
  - The resistance and capacitance of each segment are given by r L/N and c L/N

\[ \tau_{DN} = \frac{(L/N)^2}{N(N+1)/(2N^2)} = CR\left(\frac{N+1}{2N}\right) \]

where R (= rL) and C (= cL) are the total lumped resistance and capacitance of the wire

- For large N
  \[ \tau_{DN} = RC/2 = rcL^2/2 \]
  - Delay of a wire is a quadratic function of its length, L
  - The delay is 1/2 of that predicted (by the lumped model)
### Step Response Points

<table>
<thead>
<tr>
<th>Voltage Range</th>
<th>Lumped RC</th>
<th>Distributed RC</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0 \rightarrow 50% \ (t_p)$</td>
<td>$0.69 \ \text{RC}$</td>
<td>$0.38 \ \text{RC}$</td>
</tr>
<tr>
<td>$0 \rightarrow 63% \ (\tau)$</td>
<td>$\text{RC}$</td>
<td>$0.5 \ \text{RC}$</td>
</tr>
<tr>
<td>$10% \rightarrow 90% \ (t_r)$</td>
<td>$2.2 \ \text{RC}$</td>
<td>$0.9 \ \text{RC}$</td>
</tr>
<tr>
<td>$0 \rightarrow 90%$</td>
<td>$2.3 \ \text{RC}$</td>
<td>$1.0 \ \text{RC}$</td>
</tr>
</tbody>
</table>

Time to reach the $50\%$ point is $t = \ln(2) \tau = 0.69 \tau$

Time to reach the $90\%$ point is $t = \ln(9) \tau = 2.2 \tau$

- **Example:** Consider a Al1 wire 10 cm long and 1 μm wide
  
  - Using a lumped C only model with a source resistance ($R_{\text{Driver}}$) of 10 kΩ and a total lumped capacitance ($C_{\text{lumped}}$) of 11 pF
    
    $t_{50\%} = 0.69 \times 10 \text{ kΩ} \times 11\text{pF} = 76 \text{ ns}$
    
    $t_{90\%} = 2.2 \times 10 \text{ kΩ} \times 11\text{pF} = 242 \text{ ns}$

  - Using a distributed RC model with $c = 110 \text{ aF}/\mu\text{m}$ and $r = 0.075 \ \text{Ω}/\mu\text{m}$
    
    $t_{50\%} = 0.38 \times (0.075 \ \text{Ω}/\mu\text{m}) \times (110 \text{ aF}/\mu\text{m}) \times (10^5 \ \mu\text{m})^2 = 31.4 \text{ ns}$
    
    $t_{90\%} = 0.9 \times (0.075 \ \text{Ω}/\mu\text{m}) \times (110 \text{ aF}/\mu\text{m}) \times (10^5 \ \mu\text{m})^2 = 74.25 \text{ ns}$

    Poly: $t_{50\%} = 0.38 \times (150 \ \text{Ω}/\mu\text{m}) \times (88+2 \times 54 \text{ aF}/\mu\text{m}) \times (10^5 \ \mu\text{m})^2 = 112 \ \mu\text{s}$
    
    Al5: $t_{50\%} = 0.38 \times (0.0375 \ \text{Ω}/\mu\text{m}) \times (5.2+2 \times 12 \text{ aF}/\mu\text{m}) \times (10^5 \ \mu\text{m})^2 = 4.2 \text{ ns}$
Putting It All Together

- Total propagation delay consider driver and wire

\[ \tau_D = R_{\text{Driver}}C_w + \frac{(R_wC_w)}{2} = R_{\text{Driver}}C_w + 0.5r_wc_wL^2 \]

and

\[ t_p = 0.69 R_{\text{Driver}}C_w + 0.38 R_wC_w \]

where \( R_w = r_wL \) and \( C_w = c_wL \)

- The delay introduced by wire resistance becomes dominant when \( \frac{(R_wC_w)}{2} \geq R_{\text{Driver}}C_w \) (when \( L \geq 2R_{\text{Driver}}/R_w \))

  - For an \( R_{\text{Driver}} = 1 \text{ k}\Omega \) driving an 1 \( \mu \text{m} \) wide Al1 wire, \( L_{\text{crit}} \) is 2.67 cm
rc delays should be considered when $t_{pRC} > t_{pgate}$ of the driving gate

$$L_{crit} > \sqrt{\frac{t_{pgate}}{0.38rc}}$$

- actual $L_{crit}$ depends upon the size of the driving gate and the interconnect material

rc delays should be considered when the rise (fall) time at the line input is smaller than RC, the rise (fall) time of the line

$$t_{rise} < RC$$

- when not met, the change in the signal is slower than the propagation delay of the wire so a lumped C model suffices
Overcoming Interconnect Resistance

- Selective technology scaling
  - scale W while holding H constant

- Use better interconnect materials
  - lower resistivity materials like copper
    - As processes shrink, wires get shorter (reducing C) but they get closer together (increasing C) and narrower (increasing R). So RC wire delay increases and capacitive coupling gets worse.
    - Copper has about 40% lower resistivity than aluminum, so copper wires can be thinner (reducing C) without increasing R
  - use silicides (WSi₂, TiSi₂, PtSi₂ and TaSi)
    - Conductivity is 8-10 times better than poly alone

- Use more interconnect layers
  - reduces the average wire length L (but beware of extra contacts)
Wire Spacing Comparisons

Intel P856.5
Al, 0.25\(\mu\)m

Ω - 0.05
Ω - 0.12
Ω - 0.33
Ω - 0.33
Ω - 1.11

Scale: 2,160 nm

Intel P858
Al, 0.18\(\mu\)m

Ω - 0.07
Ω - 0.08
Ω - 0.17
Ω - 0.49
Ω - 1.00

IBM CMOS-8S
CU, 0.18\(\mu\)m

Ω - 0.10
Ω - 0.10
Ω - 0.50
Ω - 0.50
Ω - 0.70
Ω - 0.97

From MPR, 2000
Comparison of Wire Delays

![Comparison of Wire Delays](image)

From MPR, 2000