Chapter 4

Electrical Characteristics of CMOS

Jin-Fu Li
Department of Electrical Engineering
National Central University
Jungli, Taiwan
Outline

- Resistance & Capacitance Estimation
- DC Response
- Logic Level and Noise Margins
- Transient Response
- Delay Estimation
- Transistor Sizing
- Power Analysis
- Scaling Theory
Resistance Estimation

- **Resistance**
  - \( R = (\rho / t)(L / W) \), where \((\rho, t, L, W)\) is (resistivity, thickness, conductor length, conductor width)

- **Sheet resistance**
  - \( R_s = \Omega / \square \)
  - Thus \( R = R_s (L / W) \)

1 rectangular block
\[ R = R_s (L / W) \]

4 rectangular block
\[ R = R_s (2L / 2W) = R_s (L / W) \]
Drain-Source MOS Resistance

- A simplified linear model of MOS is useful at the logic level design
  - RC model of an NMOS
  - The drain-source resistance at any point on the current curve as shown below

$$I_{ds} \quad V_{ds}$$
Drain-Source Resistance

- The resistance at point a
  - The current is approximated by
    - $I_{ds} \approx \beta_n (V_{gs} - V_t)V_{ds}$
  - Thus the resistance is
    - $R_n \approx 1/ \beta_n (V_{gs} - V_t)$

- The resistance at point b
  - The full non-saturated current must be used so that
    - $I_{ds} = \frac{1}{2} \beta_n [2(V_{gs} - V_t)V_{ds} - V_{ds^2}]$
  - Thus the resistance is
    - $R_n = 2 / \beta_n [2(V_{gs} - V_t) - V_{ds}]$
Drain-Source Resistance

- The resistance at point c
  - The current is
    - \( I_{ds} \approx \frac{1}{2} \beta_n (V_{gs} - V_i)^2 \)
  - Thus the resistance is
    - \( R_n = 2V_{ds} / \beta_n (V_{gs} - V_i)^2 \)
  - \( R_n \) is a function of both \( V_{gs} \) and \( V_{ds} \)

- These equations show that it is not possible to define a constant value for \( R_n \)

- However, \( R_n \) is inversely proportion to \( \beta_n \) in all cases, i.e.,
  - \( R_n \propto 1 / \beta_n \)
  - \( \beta_n = k(W/L) \), \( W/L \) is called aspect ratio
Capacitance Estimation

- The switching speed of MOS circuits are heavily affected by the parasitic capacitances associated with the MOS device and interconnection capacitances.
- The total load capacitance on the output of a CMOS gate is the sum of:
  - Gate capacitance
  - Diffusion capacitance
  - Routing capacitance
- Understanding the source of parasitic loads and their variations is essential in the design process.
MOS-Capacitor Characteristics

- The capacitance of an MOS is varied with the applied voltages.

- Capacitance can be calculated by
  
  \[ C = \frac{\varepsilon_0 \varepsilon_x A}{d} \]

  - \( \varepsilon_x \) is dielectric constant
  - \( \varepsilon_0 \) is permittivity of free space

- Depend on the gate voltage, the state of the MOS surface may be in
  
  - Accumulation
  - Depletion
  - Inversion
MOS Capacitor Characteristics

- When $V_g < 0$, an accumulation layer is formed
  - The negative charge on the gate attracts holes toward the silicon surface
  - The MOS structure behaves like a parallel-plate capacitor

\[ C_0 = \frac{\varepsilon_0 \varepsilon_{SiO_2}}{t_{ox}} A \]
MOS Capacitor Characteristics

- When a small positive voltage is applied to the gate, a depletion layer is formed.
- The positive gate voltage repels holes, leaving a negatively charged region depleted of carriers.

\[ C_{dep} = \frac{\varepsilon_0 \varepsilon_{Si}}{d} A \]

\[ C_{gb} = \frac{C_0 C_{dep}}{C_0 + C_{dep}} \]
MOS Capacitor Characteristics

- When the gate voltage is further increased, an n-type channel (inversion layer) is created.
- If the MOS is operated at high frequency, the surface charge is not able to track fast moving gate voltages.

\[ C_{gb} = C_0 \]  

\[ C_{gb} = \frac{C_0 C_{dep}}{C_0 + C_{dep}} = C_{min} \]
Consequently, the dynamic gate capacitance as a function of gate voltage, as shown below

- The minimum capacitance depends on the depth of the depletion region, which depends on the substrate doping density.
MOS Device Capacitances

- The parasitic capacitances of an MOS transistor are shown as below
  - $C_{gs}$, $C_{gd}$: gate-to-channel capacitances, which are lumped at the source and the drain regions of the channel, respectively
  - $C_{sb}$, $C_{db}$: source and drain-diffusion capacitances to bulk
  - $C_{gb}$: gate-to-bulk capacitance

\[ C_g = C_{gb} + C_{gs} + C_{gd} \]
Variation of Gate Capacitance

The behavior of the gate capacitance in the three regions of operation is summarized as below

- **Off region** ($V_{gs} < V_t$): $C_{gs} = C_{gd} = 0$; $C_g = C_{gb}$
- **Non-saturated region** ($V_{gs} - V_t > V_{ds}$): $C_{gs}$ and $C_{gd}$ become significant. These capacitances are dependent on gate voltage. Their value can be estimated as

$$C_{gd} = C_{gs} = \frac{1}{2} \frac{\varepsilon_0 \varepsilon_{SiO_2}}{t_{ox}} A$$

- **Saturated region** ($V_{gs} - V_t < V_{ds}$): The drain region is pinched off, causing $C_{gd}$ to be zero. $C_{gs}$ increases to approximately

$$C_{gs} = \frac{2}{3} \frac{\varepsilon_0 \varepsilon_{SiO_2}}{t_{ox}} A$$
Approximation of the $C_g$

☐ The $C_g$ can be further approximated with

\[ C_g = C_{ox} A, \text{ where } C_{ox} = \frac{\varepsilon_o \varepsilon_{SiO_2}}{t_{ox}} \]

☐ The gate capacitance is determined by the gate area, since the thickness of oxide is associated with process of fabrication

☐ For example, assume that the thickness of silicon oxide of the given process is $150 \times 10^{-8} \mu m$. Calculate the capacitance of the MOS shown below

\[ C_g = \frac{3.9 \times 8.854 \times 10^{-14}}{150 \times 10^{-8}} \times 2 = 25.5 \times 2 \times 10^{-4} \text{ pF} \approx 0.005 \text{ pF} \]
Diffusion Capacitance

- Diffusion capacitance $C_d$ is proportional to the diffusion-to-substrate junction area

\[ C_d = C_{ja} \times (ab) + C_{jp} \times (2a + 2b) \]

- $C_{ja}$ = junction capacitance per micron square
- $C_{jp}$ = periphery capacitance per micron
Junction Capacitance

- Semiconductor physics reveals that a PN junction automatically exhibits capacitance due to the opposite polarity charges involved. This is called junction or depletion capacitance and is found at every drain or source region of a MOS.

- The junction capacitance is varies with the junction voltage, it can be estimate as

\[ C_j = C_{j0} \left(1 - \frac{V_j}{V_b}\right)^{-m} \]

- \( C_j \): junction voltage (negative for reverse bias)
- \( C_{j0} \): zero bias junction capacitance (\( V_j = 0 \))
- \( V_b \): built-in junction voltage \( \sim 0.6V \)
Single Wire Capacitance

- Routing capacitance between metal and substrate can be approximated using a parallel-plate model.

![Diagram of parallel-plate model with fringing fields and dimensions W, L, T, H, and substrate and insulator (Oxide).]

- In addition, a conductor can exhibit capacitance to an adjacent conductor on the same layer.
Multiple Conductor Capacitances

- Modern CMOS processes have multiple routing layers
  - The capacitance interactions between layers can become quite complex

- Multilevel-layer capacitance can be modeled as below

\[
C_2 = C_{21} + C_{23} + C_{22}
\]
A Process Cross Section

- Interlayer capacitances of a two-level-metal process

Diagram showing a cross-sectional view of a multi-layered metal process with labeled components such as m1, m2, poly, and C, with the substrate labeled at the bottom.
Inductor

- For bond wire inductance
  \[ L = \frac{\mu}{2\pi} \ln\left(\frac{4h}{d}\right) \]

- For on-chip metal wires
  \[ L = \frac{\mu}{2\pi} \ln\left(\frac{8h + w}{w + 4h}\right) \]

- The inductance produces \( L\frac{di}{dt} \) noise especially for ground bouncing noise. Note that when CMOS circuit are clocked, the current flow changes greatly.

\[ V = L \frac{di}{dt} \]
Distributed RC Effects

- The propagation delay of a signal along a wire mainly depends on the distributed resistance and capacitance of the wire.
- A long wire can be represented in terms of several RC sessions, as shown below.

\[
\begin{align*}
\text{The response at node } V_j \text{ with respect to time is then given by:}
\end{align*}
\]

\[
CdV = Idt \Rightarrow C \frac{dV_j}{dt} = (I_{j-1} - I_j) = \frac{(V_{j-1} - V_j)}{R} - \frac{(V_j - V_{j+1})}{R}
\]
Distributed RC Effects

- As the number of sections in the network becomes large (and the sections become small), the above expression reduces to the differential form:
  \[ rc \frac{dV}{dt} = \frac{d^2V}{dx^2} \Rightarrow t_x = kx^2 \]
  - \( r \) : resistance per unit length
  - \( c \) : capacitance per unit length

- Alternatively, a discrete analysis of the circuit shown in the previous page yields an approximate signal delay of:
  \[ t_n = 0.7 \times \frac{RCn(n+1)}{2} \]
  , where \( n \)=number of sections
  \[ t_1 = 0.7 \frac{rcl^2}{2} \]
Wire Segmentation with Buffers

- To optimize speed of a long wire, one effective method is to segment the wire into several sections and insert buffers within these sections.

- Consider a poly bus of length 2mm that has been divided into two 1mm sections.
  - Assume that \( t_x = 4 \times 10^{-15} x^2 \)
  - With buffer \( t_p = 4 \times 10^{-15} \times 1000^2 + t_{buf} + 4 \times 10^{-15} \times 1000^2 = 4ns + t_{buf} + 4ns = 8ns + t_{buf} \)
  - Without buffer \( t_p = 4 \times 10^{-15} \times 2000^2 = 16ns \)

- By keeping the buffer delay small, significant gain can be obtained with buffer insertion.
Crosstalk

- A capacitor does not like to change its voltage instantaneously.
- A wire has high capacitance to its neighbor.
  - When the neighbor switches from 1→0 or 0→1, the wire tends to switch too.
  - Called capacitive coupling or crosstalk.
- Crosstalk effects
  - Noise on nonswitching wires
  - Increased delay on switching wires
Crosstalk Delay

- Assume layers above and below on average are quiet
  - Second terminal of capacitor can be ignored
  - Model as $C_{gnd} = C_{top} + C_{bot}$
- Effective $C_{adj}$ depends on behavior of neighbors
  - Miller effect

<table>
<thead>
<tr>
<th>B</th>
<th>$\Delta V$</th>
<th>$C_{eff(A)}$</th>
<th>MCF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>$V_{DD}$</td>
<td>$C_{gnd} + C_{adj}$</td>
<td>1</td>
</tr>
<tr>
<td>Switching with A</td>
<td>0</td>
<td>$C_{gnd}$</td>
<td>0</td>
</tr>
<tr>
<td>Switching opposite A</td>
<td>$2V_{DD}$</td>
<td>$C_{gnd} + 2C_{adj}$</td>
<td>2</td>
</tr>
</tbody>
</table>
Crosstalk Noise

- Crosstalk causes noise on nonswitching wires
- If victim is floating:
  - model as capacitive voltage divider

$$\Delta V_{\text{victim}} = \frac{C_{\text{adj}}}{C_{\text{gnd-v}} + C_{\text{adj}}} \Delta V_{\text{aggressor}}$$

![Crosstalk Noise Diagram]

*Advanced Reliable Systems (ARES) Lab.*  
*Jin-Fu Li, EE, NCU*
Driven Victim

- Usually victim is driven by a gate that fights noise
- Noise depends on relative resistances
- Victim driver is in linear region, agg. in saturation
- If sizes are same, $R_{aggressor} = 2-4 \times R_{victim}$

\[
\Delta V_{victim} = \frac{C_{adj}}{C_{gnd-v} + C_{adj}} \frac{1}{1 + k} \Delta V_{aggressor}
\]

\[
k = \frac{\tau_{aggressor}}{\tau_{victim}} = \frac{R_{aggressor}}{R_{victim}} \left( \frac{C_{gnd-a} + C_{adj}}{C_{gnd-v} + C_{adj}} \right)
\]
Simulation Waveforms

Simulated coupling for $C_{adj} = C_{victim}$

0.3 0.6 0.9 1.2 1.5 1.8
Victim (undriven): 50%
Victim (half size driver): 16%
Victim (equal size driver): 8%
Victim (double size driver): 4%

$0 200 400 600 800 1000 1200 1400 1800 2000$

$\textbf{t(ps)}$
DC Response

- DC Response: $V_{out}$ vs. $V_{in}$ for a gate
- Ex: Inverter
  - When $V_{in} = 0 \rightarrow V_{out} = V_{DD}$
  - When $V_{in} = V_{DD} \rightarrow V_{out} = 0$
  - In between, $V_{out}$ depends on transistor size and current $I_{dsp}$
  - By KCL, must settle such that $I_{dsn} = |I_{dsp}|$
  - We could solve equations
  - But graphical solution gives more insight
Transistor Operation

- Current depends on region of transistor behavior
- For what $V_{\text{in}}$ and $V_{\text{out}}$ are NMOS and PMOS in
  - Cutoff?
  - Linear?
  - Saturation?
**NMOS Operation**

<table>
<thead>
<tr>
<th>Cutoff</th>
<th>Linear</th>
<th>Saturated</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_{gsn} &lt; V_{tn}$</td>
<td>$V_{gsn} &gt; V_{tn}$</td>
<td>$V_{gsn} &gt; V_{tn}$</td>
</tr>
<tr>
<td></td>
<td>$V_{dsn} &lt; V_{gsn} - V_{tn}$</td>
<td>$V_{dsn} &gt; V_{gsn} - V_{tn}$</td>
</tr>
</tbody>
</table>

$V_{gsn} = V_{in}$  
$V_{dsn} = V_{out}$

Advanced Reliable Systems (ARES) Lab.  
Jin-Fu Li, EE, NCU
## NMOS Operation

<table>
<thead>
<tr>
<th>Cutoff</th>
<th>Linear</th>
<th>Saturated</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_{gsn} &lt; V_{tn}$</td>
<td>$V_{gsn} &gt; V_{tn}$</td>
<td>$V_{gsn} &gt; V_{tn}$</td>
</tr>
<tr>
<td>$V_{in} &lt; V_{tn}$</td>
<td>$V_{in} &gt; V_{tn}$</td>
<td>$V_{in} &gt; V_{tn}$</td>
</tr>
<tr>
<td></td>
<td>$V_{dsn} &lt; V_{gsn} - V_{tn}$</td>
<td>$V_{dsn} &gt; V_{gsn} - V_{tn}$</td>
</tr>
<tr>
<td></td>
<td>$V_{out} &lt; V_{in} - V_{tn}$</td>
<td>$V_{out} &gt; V_{in} - V_{tn}$</td>
</tr>
</tbody>
</table>

**Equations:**

- $V_{gsn} = V_{in}$
- $V_{dsn} = V_{out}$

**Diagram:**

![NMOS Circuit Diagram]
# PMOS Operation

<table>
<thead>
<tr>
<th>Cutoff</th>
<th>Linear</th>
<th>Saturated</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_{gsp} &gt; V_{tp}$</td>
<td>$V_{gsp} &lt; V_{tp}$</td>
<td>$V_{gsp} &lt; V_{tp}$</td>
</tr>
<tr>
<td>$V_{dsp} &gt; V_{gsp} - V_{tp}$</td>
<td></td>
<td>$V_{dsp} &lt; V_{gsp} - V_{tp}$</td>
</tr>
</tbody>
</table>

\[
V_{gsp} = V_{in} - V_{DD} \quad V_{tp} < 0 \\
V_{dsp} = V_{out} - V_{DD}
\]
### PMOS Operation

<table>
<thead>
<tr>
<th>Cutoff</th>
<th>Linear</th>
<th>Saturated</th>
</tr>
</thead>
</table>
| $V_{gsp} > V_{tp}$  
$V_{in} > V_{DD} + V_{tp}$ | $V_{gsp} < V_{tp}$  
$V_{in} < V_{DD} + V_{tp}$  
$V_{dsp} > V_{gsp} - V_{tp}$  
$V_{out} > V_{in} - V_{tp}$ | $V_{gsp} < V_{tp}$  
$V_{in} < V_{DD} + V_{tp}$  
$V_{dsp} < V_{gsp} - V_{tp}$  
$V_{out} < V_{in} - V_{tp}$ |

$$V_{gsp} = V_{in} - V_{DD}$$  
$$V_{tp} < 0$$  
$$V_{dsp} = V_{out} - V_{DD}$$

---

Advanced Reliable Systems (ARES) Lab.  
Jin-Fu Li, EE, NCU
Make pMOS is wider than nMOS such that $\beta_n = \beta_p$
Current & $V_{\text{out}}$, $V_{\text{in}}$
Load Line Analysis

- For a given $V_{in}$:
  - Plot $I_{dsn}$, $I_{dsp}$ vs. $V_{out}$
  - $V_{out}$ must be where $|currents|$ are equal in

$V_{in}$

$V_{in1}$

$V_{in2}$

$V_{in3}$

$V_{in4}$

$V_{in5}$

$V_{out}$

$V_{DD}$
DC Transfer Curve

Transcribe points onto $V_{in}$ vs. $V_{out}$ plot.
Operation Regions

- Revisit transistor operating regions

<table>
<thead>
<tr>
<th>Region</th>
<th>nMOS</th>
<th>pMOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Cutoff</td>
<td>Linear</td>
</tr>
<tr>
<td>B</td>
<td>Saturation</td>
<td>Linear</td>
</tr>
<tr>
<td>C</td>
<td>Saturation</td>
<td>Saturation</td>
</tr>
<tr>
<td>D</td>
<td>Linear</td>
<td>Saturation</td>
</tr>
<tr>
<td>E</td>
<td>Linear</td>
<td>Cutoff</td>
</tr>
</tbody>
</table>

Advanced Reliable Systems (ARES) Lab.  Jin-Fu Li, EE, NCU
If $\beta_p / \beta_n \neq 1$, switching point will move from $V_{DD}/2$.

Called *skewed gate*.

Other gates: collapse into equivalent inverter.

\[ \beta_p = 0.1, \quad \beta_n = 0.5, \quad \beta_p / \beta_n = 10, \quad \beta_p / \beta_n = 0.1 \]
**Noise Margin**

- How much noise can a gate input see before it does not recognize the input?

```
<table>
<thead>
<tr>
<th>Input Characteristics</th>
<th>Output Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logical High</td>
<td>V_{DD}</td>
</tr>
<tr>
<td>Logical Low</td>
<td>GND</td>
</tr>
</tbody>
</table>
```

- **Logical High Output Range**
  - NM_H
  - V_{OH}
  - V_{IH}

- **Logical Low Output Range**
  - NM_L
  - V_{OL}
  - V_{IL}

- **Indeterminate Region**
  - Logical High Input Range
  - Logical Low Input Range

---

Advanced Reliable Systems (ARES) Lab.

Jin-Fu Li, EE, NCU
Transient Analysis

- *DC analysis* tells us $V_{out}$ if $V_{in}$ is constant
- *Transient analysis* tells us $V_{out}(t)$ if $V_{in}(t)$ changes
Switching Characteristics

Switching characteristics for CMOS inverter

\[ V_{ds} = V_{gs} - V_t \]

\[ V_{in}(t) \quad V_{out}(t) \]

\[ V_{DD} \quad V_{in}(t) \quad V_{out}(t) \]

\[ t_{df} \quad t_{dr} \]

90% 50% 10%
Switching Characteristics

- **Rise time \( (t_r) \)**
  - The time for a waveform to rise from 10% to 90% of its steady-state value

- **Fall time \( (t_f) \)**
  - The time for a waveform to fall from 90% to 10% steady-state value

- **Delay time \( (t_d) \)**
  - The time difference between input transition (50%) and the 50% output level. (This is the time taken for a logic transition to pass from input to output)
    - **High-to-low delay \( (t_{df}) \)**
    - **Low-to-high delay \( (t_{dr}) \)**
Fall Time of the Inverter

- Equivalent circuit for fall-time analysis

\[
\text{Input rising}
\]

- The fall time consists of two intervals
  - \( t_{f1} \): period during which the capacitor voltage, \( V_{\text{out}} \), drops from 0.9\( V_{\text{DD}} \) to \( (V_{\text{DD}} - V_{\text{tn}}) \)
  - \( t_{f2} \): period during which the capacitor voltage, \( V_{\text{out}} \), drops from \( (V_{\text{DD}} - V_{\text{tn}}) \) to 0.1\( V_{\text{DD}} \)
Timing Calculation

- $t_{f1}$ can be calculated with the current-voltage equation as shown below, while in saturation:
  \[
  C_L \frac{dV_{out}}{dt} + \frac{\beta_n}{2} (V_{DD} - V_m)^2 = 0
  \]
- $t_{f2}$ also can be obtained by the same way.
- Finally, the fall time can be estimated with:
  \[
  t_f \approx k \times \frac{C_L}{\beta_n V_{DD}}
  \]

- Similarly, the rise time can be estimated with:
  \[
  t_r \approx k \times \frac{C_L}{\beta_p V_{DD}}
  \]

- Thus the propagation delay is:
  \[
  t_p \approx k \times \frac{C_L}{V_{DD}} \left( \frac{1}{\beta_n} + \frac{1}{\beta_p} \right)
  \]
Design Challenges

☐ \( \beta_n = \beta_p \), rise time=fall time
  - This implies \( W_p = 2-3W_n \)

☐ Reduce \( C_L \)
  - Careful layout can help to reduce the diffusion and interconnect capacitance

☐ Increase \( \beta_n \) and \( \beta_p \)
  - Increase the transistor sizes also increases the diffusion capacitance as well as the gate capacitance. The latter will increase the fan-out factor of the driving gate and adversely affect its speed

☐ Increase \( V_{DD} \)
  - Designers don’t have too much control over this
Gate Delays

- Consider a 3-input NAND gate as shown below

- When pull-down path is conducting
  - \( \beta_{\text{neff}} = \frac{1}{(1/\beta_{n1}) + (1/\beta_{n2}) + (1/\beta_{n3})} \)
  - For \( \beta_{n1} = \beta_{n2} = \beta_{n3} \Rightarrow \beta_{\text{neff}} = \frac{\beta_n}{3} \)

- When the pull-down path is conducting
  - Only one p-transistor has to turn on to raise the output.
    Thus \( \beta_{\text{peff}} = \beta_p \)
Gate Delays

- Graphical illustration of the effect of series transistors

- In general, the fall time $t_f$ is $mt_f (t_f/m)$ for $m$ n-transistors in series (parallel). Similarly, the rise time $t_r$ for $k$ p-transistors in series (parallel) is $kt_r (t_r/k)$
Switch-Level RC Model

- **RC modeling**
  - Transistors are regarded as a resistance discharging or charging a capacitance

- **Simple RC modeling**
  - Lumped RCs
    - \[ t_{df} = \sum R_{pulldown} \times \sum C_{pulldown-path} \]
  - Elmore RC modeling
    - Distributed RCs
    - \[ t_d = \sum_i R_i C_i \]
Consider a 4-input NAND as shown below

**Simple RC model**

\[ t_{df} = \sum R_{\text{pulldown}} \times \sum C_{\text{pulldown-path}} \]

\[ = (R_{N1} + R_{N2} + R_{N3} + R_{N4}) \times (C_{\text{out}} + C_{ab} + C_{bc} + C_{cd}) \]

\[ t_{dr} = R_{p4} \times C_{\text{out}} \]

**Elmore RC model**

\[ t_{d} = \sum R_{i} C_{i} \]

\[ t_{df} = (R_{N1} \times C_{cd}) + [(R_{N1} + R_{N2}) \times C_{bc}] \]

\[ + [(R_{N1} + R_{N2} + R_{N3}) \times C_{ab}] \]

\[ + [(R_{N1} + R_{N2} + R_{N3} + R_{N4}) \times C_{\text{out}}] \]
Cascaded CMOS Inverter

- As discussed above, if we want to have approximately the same rise and fall times for an inverter, for current CMOS process, we must make
  - \( W_p = 2-3W_n \)
  - Increase layout area and dynamic power dissipation

- In some cascaded structures it is possible to use minimum or equal-size devices without compromising the switching response

- In the following, we illustrate two examples to explain why it is possible
Cascaded CMOS Inverter

Example 1:

\[
I_{\text{charge}} = \frac{4}{1}
\]
\[
R \quad 3C_{\text{eq}}
\]
\[
I_{\text{discharge}} = \frac{2}{1}
\]
\[
W_p = 2W_n
\]

\[
t_{\text{inv-pair}} = t_{\text{fall}} + t_{\text{rise}}
\]
\[
= R \cdot 3C_{\text{eq}} + \frac{R}{2} \cdot 3C_{\text{eq}}
\]
\[
= 3RC_{\text{eq}} + 3RC_{\text{eq}}
\]
\[
= 6RC_{\text{eq}}
\]

Example 2:

\[
I_{\text{charge}} = \frac{2R}{1}
\]
\[
R \quad 2C_{\text{eq}}
\]
\[
I_{\text{discharge}} = \frac{2C_{\text{eq}}}{1}
\]
\[
W_p = W_n
\]

\[
t_{\text{inv-pair}} = t_{\text{fall}} + t_{\text{rise}}
\]
\[
= R \cdot 2C_{\text{eq}} + 2R \cdot 2C_{\text{eq}}
\]
\[
= 6RC_{\text{eq}}
\]
Stage Ratio

- To drive large capacitances such as long buses, I/O buffers, etc.
  - Using a chain of inverters where each successive inverter is made larger than the previous one until the last inverter in the chain can drive the large load in the time required
  - The ratio by which each stage is increased in size is called stage ratio

- Consider the circuit shown below
  - It consists of n-cascaded inverters with stage-ratio $a$ driving a capacitance $C_L$

\[ \text{Stage Ratio} \]

\[ \text{To drive large capacitances such as long buses, I/O buffers, etc.} \]

\[ \text{Using a chain of inverters where each successive inverter is made larger than the previous one until the last inverter in the chain can drive the large load in the time required} \]

\[ \text{The ratio by which each stage is increased in size is called stage ratio} \]

\[ \text{Consider the circuit shown below} \]

\[ \text{It consists of n-cascaded inverters with stage-ratio $a$ driving a capacitance $C_L$} \]
Stage Ratio

- The delay through each stage is $a t_d$, where $t_d$ is the average delay of a minimum-sized inverter driving another minimum-sized inverter.
- Hence the delay through $n$ stages is $n a t_d$.
- If the ratio of the load capacitance to the capacitance of a minimum inverter, $C_L/C_g$, is $R$, then $a^n = R$.
  - Hence $\ln(R) = n \ln(a)$.
  - Thus the total delay is $\ln(R) (a/\ln(a)) t_d$.
  - The optimal stage ratio may be determined from

$$a_{opt} = e^{\frac{k + a_{opt}}{a_{opt}}}$$

where $k$ is $\frac{C_{\text{drain}}}{C_{\text{gate}}}$.
Power Dissipation

- **Instantaneous power**
  - The value of power consumed at any given instant
  - \( P(t) = v(t)i(t) \)

- **Peak power**
  - The highest power value at any given instant; peak power determines the component’s thermal and electrical limits and system packaging requirements
  - \( P_{\text{peak}} = Vi_{\text{peak}} \)

- **Average power**
  - The total distribution of power over a time period; average power impacts the battery lifetime and heat dissipation
  - \( P_{\text{ave}} = \frac{1}{T} \int_{t}^{t+T} P(t) \, dt = \frac{V}{T} \int_{t}^{t+T} i(t) \, dt \)
Power Analysis for CMOS Circuits

- Two components of power consumption in a CMOS circuit
  - Static power dissipation
    - Caused by the leakage current and other static current
  - Dynamic power dissipation
    - Caused by the total output capacitance
    - Caused by the short-circuit current

- The total power consumption of a CMOS circuit is

\[ P_t = P_s + P_{sw} + P_{sc} \]

- \( P_s \): static power (leakage power)
- \( P_{sw} \): switching power
- \( P_{sc} \): short-circuit power

Advanced Reliable Systems (ARES) Lab.

Jin-Fu Li, EE, NCU
Static Power

- Static dissipation is major contributed by
  - Reverse bias leakage between diffusion regions and the substrate
  - Subthreshold conduction

\[
i_0 = i_s \left( e^{qV/KT} - 1 \right)
\]

\[
P_s = \sum_{1}^{n} I_{leakage} \times V_{supply}
\]

n=number of devices
Dynamic Power Dissipation

- **Switching power**
  - Caused by charging and discharging the output capacitive load

- **Consider an inverter operated at a switching frequency** \( f = 1/T \)

\[
P_{sw} = \frac{1}{T} \int_0^T i_o(t) v_o(t) dt
\]

\[
i_p = i_o = C_L \frac{dv_o}{dt}
\]

\[
i_n = -i_o = -C_L \frac{dv_o}{dt}
\]

\[
P_{sw} = \frac{1}{T} \left[ \int_0^{V_{DD}} C_L v_o dv_o - \int_0^0 C_L v_o dv_o \right]
\]

\[
P_{sw} = \frac{C_L V_{DD}^2}{T} = fC_L V_{DD}^2
\]
Power & Energy

- Energy consumption of an inverter (from $0 \rightarrow V_{DD}$)
  - The energy drawn from the power supply is
    - $E = QV = C_L V_{DD}^2$
  - The energy stored in the load capacitance is
    - $E_{\text{cap}} = \int_0^{V_{DD}} C_v o \, dV_o = \frac{1}{2} C_L V_{DD}^2$
  - The output from $V_{DD} \rightarrow 0$
    - The $E_{\text{cap}}$ is consumed by the pull-down NMOS

- Low-energy design is more important than low-power design
  - Minimize the product of power and delay
Short-Circuit Power Dissipation

- Even if there were no load capacitance on the output of the inverter and the parasitics are negligible, the gate still dissipate switching energy.

- If the input changes slowly, both the NMOS and PMOS transistors are ON, an excess power is dissipated due to the short-circuit current.

- We are assuming that the rise time of the input is equal to the fall time.

- The short-circuit power is estimated as:
  \[ P_{sc} = I_{\text{mean}} V_{DD} \]
Short-Circuit Power Dissipation

\[ I_{\text{mean}} \text{ can be estimated as follows} \]

\[
I_{\text{mean}} = 2 \times \frac{1}{T} \left[ \int_{t_1}^{t_2} i(t) \, dt + \int_{t_2}^{t_3} i(t) \, dt \right]
\]

\[
I_{\text{mean}} = \frac{4}{T} \left[ \int_{t_1}^{t_2} i(t) \, dt \right]
\]
The NMOS transistor is operating in saturation, hence the above equation becomes

\[ I_{\text{mean}} = \frac{4}{T} \left[ \int_{t_1}^{t_2} \frac{\beta}{2} (V_{\text{in}}(t) - V_T)^2 \, dt \right] \]

\[ V_{\text{in}}(t) = \frac{V_{DD}}{t_r} t \]

\[ t_1 = \frac{V_T}{V_{DD}} t_r \]

\[ t_2 = \frac{t_r}{2} \]

\[ P_{sc} = \frac{\beta}{12} (V_{DD} - 2V_T)^3 \tau f \quad (t_r = t_f = \tau) \]
Power Analysis for Complex Gates

- The dynamic power for a complex gate cannot be estimated by the simple expression $C_L V_{DD} f$
- Dynamic power dissipation in a complex gate
  - Internal cell power
  - Capacitive load power
- Capacitive load power
  - $P_L = \alpha C_L V_{DD}^2 f$
- Internal cell power
  - $P_{int} = \sum_{i=1}^{n} \alpha_i C_i V_i V_{DD} f$
Glitch Power Dissipation

In a static logic gate, the output or internal nodes can switch before the correct logic value is being stable. This phenomenon results in spurious transitions called glitches.
Rules for Avoiding Glitch Power

- Balance delay paths; particularly on highly loaded nodes

- Insert, if possible, buffers to equalize the fast path

- Avoid if possible the cascaded design

- Redesign the logic when the power due to the glitches is an important component
Principles for Power Reduction

- Switching power dissipation
  \[ P_L = \alpha C_L V_{DD}^2 f \]
  \[ P_{int} = \sum_{i=1}^{n} \alpha_i C_i V_i V_{DD} f \]

- Prime choice: reduce voltage
  - Recent years have seen an acceleration in supply voltage reduction
  - Design at very low voltage still open question
    (0.6V...0.9V by 2010)

- Reduce switching activity
- Reduce physical capacitance
Layout Guidelines for LP Designs

- Identify, in your circuit, the high switching nodes.
- Keep the wires of high activity nodes short.
- Use low-capacitance layers (e.g., metal2, metal3, etc.) for high capacitive nodes and busses.
- Avoid, if possible, the use of dynamic logic design style.
- For any logic design, reduce the switching activity by logic reordering and balanced delays through gate tree to avoid glitch problem.
- In non-critical paths, use minimum size devices whenever it is possible without degrading the overall performance requirements.
- If pass-transistor logic style is used, careful design should be considered.
Sizing Routing Conductors

- Why do metal lines have to be sized?
  - Electromigration
  - Power supply noise and integrity (i.e., satisfactory power and signal voltage levels are presented to each gate)
  - RC delay

- Electromigration is affected by
  - Current density
  - Temperature
  - Crystal structure

- For example, the limiting value for 1 um-thick aluminum is $J_{Al} = 1 \rightarrow 2mA/\mu m$
Power & Ground Bounce

An example of ground bounce

- Voltage
  - $V_{in}$
  - $V_{out}$
  - Time

- Current
  - $V_{L}$
  - Time

- Ground bounce

$V_{L} = L\frac{di}{dt}$
Approaches for Coping with $L(di/dt)$

- Multiple power and ground pins
  - Restrict the number of I/O drivers connected to a single supply pins (reduce the $di/dt$ per supply pin)

- Careful selection of the position of the power and ground pins on the package
  - Avoid locating the power and ground pins at the corners of the package (reduce the $L$)

- Increase the rise and fall times
  - Reduce the $di/dt$

- Adding decoupling capacitances on the board
  - Separate the bonding-wire inductance from the inductance of the board interconnect
Contact Replication

- Current tends to concentrate around the perimeter in a contact hole
  - This effect, called current crowding, puts a practical upper limit on the size of the contact
  - When a contact or a via between different layers is necessary, make sure to maximize the contact perimeter (not area)
Charge Sharing

- Charge $Q = CV$
- A bus example is illustrated to explain the charge sharing phenomenon
  - A bus can be modeled as a capacitor $C_b$
  - An element attached to the bus can be modeled as a capacitor $C_s$

$$Q_b = C_b V_b$$
$$Q_s = C_s V_s$$

$$Q_T = C_b V_b + C_s V_s$$
$$V_R = \frac{Q_T}{C_T} = \frac{(C_b V_b + C_s V_s)}{(C_b + C_s)}$$

$$C_T = C_b + C_s$$

Advanced Reliable Systems (ARES) Lab.

Jin-Fu Li, EE, NCU
Design Margining

☐ The operating condition of a chip is influenced by three major factors
  - Operating temperature
  - Supply voltage
  - Process variation

☐ One must aim to design a circuit that will reliably operate over all extremes of these three variables

☐ Design corners
  - Simulating circuits at all corners is needed
    ☐ SS
    ☐ TT
    ☐ FF
Package Issues

- Packaging requirements
  - Electrical: low parasitics
  - Mechanical: reliable and robust
  - Thermal: efficient heat removal
  - Economical: cheap

- Bonding techniques

Wire Bonding

![Diagram of Wire Bonding](image-url)
Yield Estimation

\[ Y = \frac{\text{No. of good chips per wafer}}{\text{Total number of chips per wafer}} \times 100\% \]

\[ \text{Die cost} = \frac{\text{Wafer cost}}{\text{Dies per wafer} \times \text{Die yield}} \]

\[ \text{Dies per wafer} = \frac{\pi \times (\text{wafer diameter}/2)^2}{\text{die area}} - \frac{\pi \times \text{wafer diameter}}{\sqrt{2} \times \text{die area}} \]
Die Cost

Single die

Wafer

Going up to 12” (30cm)
Scaling Theory

- Consider a transistor that has a channel width $W$ and a channel length $L$
- We wish to find out how the main electrical characteristics change when both dimensions are reduced by a scaling factor $S>1$ such that the new transistor has sizes:
  - $\tilde{W} = \frac{W}{S}$
  - $\tilde{L} = \frac{L}{S}$
- Gate area of the scaled transistor:
  - $\tilde{A} = \frac{A}{S^2}$
- The aspect ratio of the scaled transistor:
  - $\frac{W}{L} = \frac{\tilde{W}}{\tilde{L}}$
Scaling Theory

- The oxide capacitance is given by
  \[ C_{ox} = \frac{e_{ox}}{t_{ox}} \]
  If the new transistor has a thinner oxide that is decreased as \( \tilde{t}_{ox} = \frac{t_{ox}}{S} \), then the scaled device has
  \[ \tilde{C}_{ox} = SC_{ox} \]

- The transconductance is increased in the scaled device to
  \[ \tilde{\beta} = S\beta \]

- The resistance is reduced in the scaled device to
  \[ \tilde{R} = \frac{1}{S\beta(V_{DD} - V_T)} = \frac{R}{S} \]
  Assume that the supply voltage is not altered
Scaling Theory

- On the other hand, if we can scale the voltages in the scaled device to the new values of
  \[ \tilde{V}_{DD} = \frac{V_{DD}}{S}, \quad \tilde{V}_T = \frac{V_T}{S} \]
  The resistance of the scaled device would be unchanged with \( \tilde{R} = R \)

- The effects of scaling the voltage, consider a scaled MOS with reduced voltages of
  \[ \tilde{V}_{DS} = \frac{V_{DS}}{S}, \quad \tilde{V}_{GS} = \frac{V_{GS}}{S} \]

- The current of the scaled device is given by
  \[ \tilde{I}_D = \frac{S\beta}{2} \left[ \frac{V_{GS}}{S} - \frac{V_T}{S} \right] \frac{V_{DS}}{S} = \frac{I_D}{S} \]

- The power dissipation of the scaled device is
  \[ \tilde{P} = \tilde{V}_{DS} \tilde{I}_D = \frac{V_{DS}I_D}{S^2} = \frac{P}{S^2} \]
Summary

- We have presented models that allow us to estimate circuit timing performance, and power dissipation.
- Guidelines for low-power design have also been presented.
- The concepts of design margining were also introduced.
- The scaling theory has also introduced.