Chapter 4. Circuit Characterization and Performance Estimation

• 4.1 introduction.
  – Each layer of transistor-forming material has both a resistance and a capacitance that are fundamental components in estimating the performance of a circuit or a system. (It also has inductance but insignificant for most on-chip circuit.)
  – The issues considered in this chapter are
    • Resistance, capacitance, and inductance calculations.
    • Delay estimations
    • Determination of conductor size for power and clock distribution.
    • Power consumption
    • Charge sharing mechanism
    • Design margining
    • Reliability
    • Effects of scaling
4.2 Resistance Estimation

- The resistance of a uniform conducting slab is  

\[ R = \left( \rho \right) \left( \frac{l}{w} \right) = R_s \left( \frac{l}{w} \right) \]

where \( \rho \) = resistivity.

\( t \) = thickness.

\( l \) = conductor length.

\( w \) = conductor width.

\( R_s \) = Sheet resistance having units of \( \Omega / \text{square} \).

- According to the above formula, the two metal slabs shown in Figure 4.1 have the same resistance.

![Figure 4.1 Determination of layer resistance](image-url)
Table 4.1 shows typical sheet resistances for 0.5 μm to 1.0 μm MOS process.

<table>
<thead>
<tr>
<th>Material</th>
<th>SHEET RESISTANCE</th>
<th>Ω/SQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intermetal (metal1-metal2)</td>
<td>0.05</td>
<td>0.1</td>
</tr>
<tr>
<td>Top-metal (metal3)</td>
<td>0.03</td>
<td>0.045</td>
</tr>
<tr>
<td>Polysilicon</td>
<td>15</td>
<td>30</td>
</tr>
<tr>
<td>Silicide</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>Diffusion (n+, p+)</td>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td>Silicided diffusion</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>n-well</td>
<td>1K</td>
<td>5K</td>
</tr>
</tbody>
</table>

Note that for metal having a given thickness, the resistivity is known, while for ploy and diffusion the resistivities are significantly influenced by the concentration density of the impurities.

From the voltage-current characteristic of an MOS transistor, the channel resistance in the linear region can be approximated as 

\[ R_c = k \left( \frac{L}{W} \right) \]

where \( k = \frac{1}{\mu C_{ox}(V_{gs} - V_t)} \) derived from equation (2.14), 1000 \(< k < 30000 \ \Omega/square \) for n- and p-channel devices.

Since the mobility of the majority carriers decreases with the increase of temperature, the channel resistance is increased by approximately 0.25% per °C for temperature above 25 °C.
• 4.21 Resistance of Nonrectangular Regions
  – Figure 4.2(a) summarizes the resistance of a number of commonly encountered shapes.
  – Figure 4.2(b) shows some shapes that are commonly encountered in practice.
  – Table 4.2 represents the results of a study to calculate the resistance of the shapes shown in Figure 4.2(b) for different dimension ratio.

![Figure 4.2](image_url)

**FIGURE 4.2** Resistance of nonrectangular shapes
© IEEE 1983
• 4.2.2 Contact and Via Resistance

  Typical values for processes currently in use (0.6µm) range form 0.25 Ω to a few tens of Ωs.

• 4.3 Capacitance Estimation

  The total load capacitance on the output of a CMOS gate is the sum of

  • gate capacitance (of other inputs connected to the output of the logic gate)
  • diffusion capacitance (of the drain regions connected to the output)
  • routing capacitance (of connections between the output and other inputs)
• 4.3.1 MOS - Capacitor Characteristics

– The capacitance-voltage characteristics of an MOS capacitor (i.e. an MOS transistor without source or drain) depend on the state of the semiconductor surface.

– The surface for a p-substrate structure shown in Figure 4.3 may be

  • **accumulation** when \( V_g < 0 \) (see Figure 4.3 (a)), and the gate capacitance may be approximated by

  \[
  C_0 = \frac{\varepsilon_{SiO_2} \varepsilon_o}{t_{ox}} \cdot A,
  \]

  where

  \( A \) = area of gate.

  \( \varepsilon_{SiO_2} \) = dielectric constant (or relative permittivity of SiO\(_2\), taken as 3.9)

  \( \varepsilon_o \) = permittivity of free space.

  • **depletion** when \( V_g > 0 \) (See Figure 4.3 (b)), and the gate capacitance

  \[
  C_{gb} = \frac{C_0 C_{dep}}{C_0 + C_{dep}} < C_o , \]

  where

  \[
  C_{dep} = \left( \frac{\varepsilon_{Si}}{d} \right) A
  \]

  is the depletion capacitance, decreased when \( d \) increases.

  \( d \) = depletion layer depth, increased when gate to substrate voltage increases.

  \( \varepsilon_{Si} \) = dielectric constant of silicon, taken as 12.

  • **inversion** as \( V_g \) is further increased (see Figure 4.3 (c)). In this case, minority carries (electrons for the p-substrate) are attracted forward the silicon surface and thus created an n-channel. Surface inversion yields a relative high conductivity layer under the gate. Thus for low frequency operation (< 100Hz), \( C_{gb} = C_o \); For high frequency operation, the surface change is not able to track fast moving gate voltages

  \[
  C_{gb} = \frac{C_o C_{dep}}{C_0 + C_{dep}} < C_o
  \]
• As it can be seen, if $V_g$ increases, the total capacitance decreases.
  – Figure 4.3 (a) plots the dynamic gate capacitance as a function of gate voltage.
4.3.2 MOS Device capacitances.

Figure 4.4 is a diagramatic representation of the parasitic capacitance for an MOS transistor. It is assumed that the overlap of the gate over source/drain is equal to zero.

The following capacitance components exist:

- \( C_{gs} \) \( C_{gd} \) = gate to channel capacitances, lumped at the source and the drain regions of the channel, respectively.
- \( C_{sb} \) \( C_{db} \) = source and drain-diffusion capacitance to bulk (or substrate).
- \( C_{gb} \) = gate to bulk capacitance.

![Diagram of MOS transistor with capacitance components labeled](image)
Figure 4.5 shows a circuit model comprising parasitic capacitances and the transistor. The total gate capacitance is given by $C_g = C_{gb} + C_{gs} + C_{gd}$.

The behavior of gate capacitance of an MOS device:

1. **Off region**, where $V_{gs} < V_t$, no channel exists, hence $C_{gs} = C_{gd} = 0$. So $C_g = C_{gb}$.

2. **Non-saturated region**, where $V_{gs} - V_t > V_{ds}$. A channel exists. Thus
   
   $$C_{gd} = C_{gs} = \frac{1}{2} \frac{\varepsilon \Phi_{SO2}}{t_ox} A \quad \text{and} \quad C_{gb} = 0.$$  

3. **Saturated region**, where $V_{gs} - V_t < V_{ds}$. The channel is pinched off at the drain end. Thus
   
   $$C_{gd} = 0 \quad \text{and} \quad C_{gs} = \frac{2}{3} \frac{\varepsilon \Phi_{SO2}}{t_ox} A.$$
- The behavior of the gate capacitance is shown in Table 4.3 where $\varepsilon = \varepsilon_o \varepsilon_{SiO_2}$

<table>
<thead>
<tr>
<th>TABLE 4.3 Approximation of intrinsic MOS gate capacitance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td>$C_{gd}$</td>
</tr>
<tr>
<td>$C_{gs}$</td>
</tr>
<tr>
<td>$C_{gd}$</td>
</tr>
<tr>
<td>$C_g = C_{gd} + C_{gs} + C_{gd}$</td>
</tr>
</tbody>
</table>

- Figure 4.6 (a) shows the $C_{gs}$ and $C_{gd}$ of a long channel n-transistor ($W = 49.2 \mu m$, $L = 4.5 \mu m$).

- Figure 4.6 (b) shows the $C_{gs}$ and $C_{gd}$ of a short channel device ($L = 0.75 \mu m$). Note that $C_{gd}$ is finite, i.e., $C_{gd} > 0$. This is due to channel side fringing fields between the gate and drain.
FIGURE 4.6 Total gate capacitance of an MOS transistor as a function of $V_{gs}$ (© IEEE 1987)
For the purpose of delay calculation, the gate capacitance can be approximated by

\[ C_g = C_{ox} A = \frac{\varepsilon \varepsilon_{SiO2}}{t_{ox}} A, \text{ where } C_{ox} = \frac{\varepsilon \varepsilon_{SiO2}}{t_{ox}} \text{ is the “thin-oxide” capacitance per unit area.} \]

- With a thin-oxide thickness \( t_{ox} \) in the order of \( 100 \rightarrow 200\text{Å} \), \( C_{ox} = \frac{3.9 \times 8.854 \times 10^{-14}}{(100 \rightarrow 200) \times 10^{-8}} = 35 \rightarrow 17 \times 10^{-4} \text{pF/\(\mu\text{m}^2\).} \]

- The gate capacitance for the case shown in Figure 4.7 for \( \lambda = 0.5 \text{ \(\mu\text{m}\), } W = 2 \text{ \(\mu\text{m}\) and } L = 1 \text{ \(\mu\text{m}\) and } t_{ox} = 150 \text{Å}, \text{ is } C_{g(\text{intrinsic})} = 2 \times 25.5 \times 10^{-4} \text{pF} = 0.005 \text{pF}. \)

---

**FIGURE 4.7** Physical Layout of a unit MOS Transistor for Capacitance Estimation
4.3.3 Diffusion (source/drain) capacitance.

- Figure 4.8 shows a model for source/drain capacitance. Total drain capacitance $C_d$ is given by
  \[ C_d = C_{ja} \cdot (a \cdot b) + C_{jp} \cdot (2a + 2b), \]
  where $C_{ja}$ = junction capacitance per $\mu m^2$,
  $C_{jp}$ = periphery capacitance per $\mu m$.
  
  \[ a = \text{width of diffusion region (\(\mu m\))}. \]
  \[ b = \text{length of diffusion region (\(\mu m\))}. \]
Table 4.4 shows typical values of diffusion capacitances for both $n$- and $p$-channel devices.

<table>
<thead>
<tr>
<th>TABLE 4.4 Typical Diffusion Capacitance Values (1μ n-well Process)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>n-DEVICE (OR WIRE)</strong></td>
</tr>
<tr>
<td>$C_{ja}$</td>
</tr>
<tr>
<td>$C_{jp}$</td>
</tr>
</tbody>
</table>

Note that the above simple capacitance calculations assume zero DC bias across the junction. However, both $C_{ja}$ and $C_{jb}$ are functions of junction voltage $V_j$ due to dependence of depletion layer thickness on $V_j$. Thus the junction capacitance is 

$$C_j = C_{jo} \left(1 - \frac{V_j}{V_b}\right)^m$$

where $V_j =$ junction voltage (negative for reversed-bias)

$C_{jo}$ = zero bias capacitance (i.e., $V_j=0$)

$V_b =$ built-in junction potential ~ 0.6 volts.

$m =$ a constant dependent on the distribution of impurities. Effective value ranges from 0.3 ~ 0.5.
• 4.3.4 SPICE Modeling of MOS capacitance
  – The SPICE MOSFET (and the corresponding model) call and the MOSFET MODEL statement are shown below

```
M1 4 3 5 0 NFET W=4U L=1U AS=15P AD=15P PS=11.5U PD=11.5U
.
.
.MODEL NFET NMOS
+ TOX=200E-8
+ CGBO=200P CGSO=600P CGDO=600P
+ CJ=200U CJSW=400P MJ=0.5 MJSW=0.3 PB=0.7
+ ....
.
.
```

• \( AS \) = source area,
• \( AD \) = drain area,
• \( PS \) = source periphery,
• \( PD \) = drain periphery.
• \textit{.MODEL} signals the beginning of MOSFET model
From these data, \( C_{g(intrinsic)} = W * L * C_{ox} = 4 * 1 * 17 * 10^{-4} \text{ pF} = 0.0068 \text{ pF} \). \( C_{ox} \) is derived from \( TOX = 200E-8 \).

Extrinsic values of \( C_{gso}, C_{gdo} \) and \( C_{gbo} \) are added to \( C_{gs}, C_{gd} \) and \( C_{gb} \) to consider the fringing field from gate terminal. They are specified in SPICE MOSFET Model by \( CGSO, CGDO \) and \( CGBO \).

\( C_{gbo} \) occurs due to the polysilicon (gate) extension beyond the channel. Thus \( C_{gbo} = CGSO * L * 2 \).

\( C_{gso} \) and \( C_{gdo} \) represent the gate source/drain capacitance due to overlap in the physical structure of the transistor. Thus, \( C_{gso} = CGSO * W \) and \( C_{gdo} = CGDO * W \).

The extrinsic gate capacitance for a typical MOS transistor is
\[
C_{g(extrinsic)} = C_{gso} + C_{gdo} + C_{gbo} = W*CGSO + W*CGDO + 2 * L * CGBO.
\]
\[
= 4 * 600 * 10^{-12} * 10^{-6} + 4 * 600 * 10^{-12} * 10^{-6} + 2 * 1 * 200 * 10^{-12} * 10^{-6} = 0.0052 \text{ pF}.
\]

\( C_{g(total)} = C_{g(intrinsic)} + C_{g(extrinsic)} = 0.0068 + 0.0052 = 0.012 \text{ pF} \).
In SPICE the source/drain capacitance is calculated as follows:

\[ C_j = (\text{Area} \cdot C_J \cdot (1 + \frac{V_J}{P_B})^{-M_J} + (\text{Periphery} \cdot C_{JSW}) \cdot (1 + \frac{V_J}{P_B})^{-M_{JSW}}) \]

where \( C_J \) = the zero-bias capacitance per junction area.

\( C_{JSW} \) = the zero-bias-junction capacitance per junction periphery.

\( M_J \) = the grading coefficient of the junction bottom.

\( M_{JSW} \) = the grading coefficient of the junction sidewall.

\( V_J \) = the junction potential.

\( P_B \) = the built-in voltage (0.4 \( \mu \rightarrow 0.8 \) V)

\( \text{Area} = AS \) or \( AD \), the area of source/drain.

\( \text{Periphery} = PS \) or \( PD \), the periphery of source/drain.

At \( V_J = 2.5V \) and \( P_B = 0.7V \),

\[ C_{j \, \text{drain}} = [15 \times 10^{-12} \cdot 2 \times 10^{-4} \cdot 2.5/0.7] + [11.5 \times 10^{-6} \cdot 4 \times 10^{-10} \cdot (1 + 2.5/0.7)^{-0.3}] \]

\[ = 0.0014 \text{pF} + 0.0029 \text{pF} = 0.0043 \text{pF} = C_{j \, \text{source}} \]
4.3.5 Routing Capacitance
4.3.5.1 Single Wire Capacitance

- Routing capacitance between metal and poly layers and the substrate is shown in Figure 4.9. It consists of three components:

1. *A parallel-plate capacitance* \( C = (\varepsilon/t)A \).

2. *Fringing capacitance to substrate*.

3. *Coupling capacitance between two metal line on the same layer*.

- An empirical formula

\[
C = \varepsilon \left[ \left( \frac{W}{h} \right) + 0.77 + 1.06 \left( \frac{W}{h} \right)^{0.25} + 1.06 \left( \frac{L}{h} \right)^{0.5} \right]
\]
4.3.5.2 Multiple Conductor Capacitance

- Figure 4.11 shows a model for routing structure with multiple conductances.

**Figure 4.11** Multilevel-layer capacitance model

- The capacitance of the middle layer (conductor of interest) is divided into three components.
  1. The line-to-ground capacitance.
  2. The line-to-line capacitance.
  3. The crossover capacitance.

- Figure 4.12 shows the capacitance of middle layer 2 to ground $C_2 = C_{21} + C_{23} + C_{22}$.

**Figure 4.12** Specific capacitances in a three-layer-metal system
– Empirical formulas to compute these capacitance values are given in Weste textbook.
– Figure 4.13 shows various capacitances for a two-level-metal process.
4.3.6 Distributed RC Effects

- The propagation of a signal along a wire depends on many factors, including the distributed resistance and capacitance of the wire, the impedance of the driving source, and the load impedance.

- A long wire can be represented in terms of several RC sections, as shown in Figure 4.15.

\[
\begin{align*}
\frac{dV_j}{dt} & = (I_{j-1} - I_j) = \frac{V_{j-1} - V_j}{R} - \frac{V_j - V_{j+1}}{R}, \\
rc \frac{dV}{dt} & = \frac{d^2V}{dx^2},
\end{align*}
\]

where

- \( r = \) resistance per unit length
- \( c = \) capacitance per unit length.
The solution for the propagation of a voltage step along a wire of length \( x \) shows that the rise/fall delay
\[
t_x = k x^2,
\]
where \( k \) is a constant.

Alternatively, a discrete analysis yields a signal delay of
\[
t_n = 0.7 \frac{R C n(n+1)}{2},
\]
where \( R \) = resistance per unit length.
\( C \) = capacitance per unit length.
\( l \) = length of wire.

Figure 4.16 shows an example of using the above formula to insert a buffer on a long wire

- without a buffer, the propagation delay
  \[
  t_p = 0.7 \times 20 \times 4 \times 10^{-4} \times 2000^2 / 2 = 112 \text{ns}
  \]

- with a buffer on the middle,
  \[
  t_p = \frac{(2 \times 0.7 \times 20 \times 4 \times 10^{-4} \times 1000^2)}{2} + t_{buf} = 66 \text{ns} + t_{buf}
  \]
Note that the above calculation considers only propagating a voltage step signal along the wire.

A model for the distributed $RC$ delay, which takes driver and receiver loading into account is shown in Figure 4.17, where

$R_s$ = the output resistance of the driver.

$C_i$ = the receiver input capacitance.

$R_t$ = the lumped resistance of the wire.

$C_l$ = the lumped capacitance of the wire.

![Figure 4.17 Simple model for RC delay calculation](image)
• **4.3.7 Capacitance Design Guide**
  
  Table 4.6 gives representative capacitance value (no fringing) for a 1 $\mu$m n-well CMOS process.

<table>
<thead>
<tr>
<th>PARAMETER</th>
<th>CAPACITANCE (ATTO FARADS ($10^{-18}$/µm$^2$))</th>
<th>COMMENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_{fe0r}$</td>
<td>300</td>
<td>n-diffusion area—varies widely with process</td>
</tr>
<tr>
<td>$C_{0p}$ ($aF$/µm$^2$)</td>
<td>400</td>
<td>n-diffusion periphery—varies widely with process</td>
</tr>
<tr>
<td>$C_{0p}$</td>
<td>500</td>
<td>p-diffusion area—varies widely with process</td>
</tr>
<tr>
<td>$C_{pp}$ ($aF$/µm$^2$)</td>
<td>400</td>
<td>p-diffusion periphery—varies widely with process</td>
</tr>
<tr>
<td>$C_{gs}$</td>
<td>1800</td>
<td>Gate capacitance—increases as $t_{ox}$ thins</td>
</tr>
<tr>
<td>$C_{g}$</td>
<td>50</td>
<td>Poly-over field oxide</td>
</tr>
<tr>
<td>$C_{m1}$</td>
<td>30</td>
<td>Metal1-over field oxide</td>
</tr>
<tr>
<td>$C_{m1p}$</td>
<td>60</td>
<td>Metal1 to poly</td>
</tr>
<tr>
<td>$C_{m1d}$</td>
<td>60</td>
<td>Metal1 to diffusion</td>
</tr>
<tr>
<td>$C_{m2}$</td>
<td>20</td>
<td>Metal2 to substrate</td>
</tr>
<tr>
<td>$C_{m2m1}$</td>
<td>30</td>
<td>Metal2 to metal1</td>
</tr>
<tr>
<td>$C_{m2p}$</td>
<td>30</td>
<td>Metal2 to poly</td>
</tr>
<tr>
<td>$C_{m2d}$</td>
<td>30</td>
<td>Metal2 to diffusion</td>
</tr>
<tr>
<td>$C_{m3}$</td>
<td>10</td>
<td>Metal3 to substrate</td>
</tr>
<tr>
<td>$C_{m3m2}$</td>
<td>30</td>
<td>Metal3 to metal2</td>
</tr>
<tr>
<td>$C_{m3m1}$</td>
<td>15</td>
<td>Metal3 to metal1</td>
</tr>
<tr>
<td>$C_{m3p}$</td>
<td>12</td>
<td>Metal3 to poly</td>
</tr>
<tr>
<td>$C_{m3d}$</td>
<td>10</td>
<td>Metal3 to diffusion</td>
</tr>
</tbody>
</table>
Example

A register that fits in a data-path is 25 μm tall (the direction of repetition). A metal2 clock line runs vertically to link all registers in an n-bit register. The register has 20 μm of 1 μm metal1, 20 μm of 1 μm poly (over field), and 16 μm of 1 μm gate capacitance.

1. Calculate the per-bit clock load and the load for a 16-bit register.

2. What would be the RC delay to the register from a clock buffer using 5 mm of 1 μm metal2 (.05 Ω/sq)?

3. How wide would the clock line have to be to keep the skew below .5 ns if a register file containing 32 16-bit registers was fed with the same 5 mm metal2 wire?

1. The parasitics are as follows:
   \[ C_{m1} = 30 \times 30 = 900 \text{ aF} \]
   \[ C_p = 20 \times 50 = 1000 \text{ aF} \]
   \[ C_{pp} = 16 \times 1800 = 28,800 \text{ aF} \]
   \[ C_{reg1} = 900 + 1000 + 28,800 = .030 \text{ pF} \]
   \[ C_{reg16} = 16 \times C_{reg1} = .48 \text{ pF} \]

2. \[ R_{metal2} = 5000 \times .05 \]
   \[ = 250 \text{ ohms} \]

   Because the capacitance load is at the end of the wire, we can approximate the RC delay by adding the metal2 track capacitance to the load capacitance and performing a simple RC calculation.

   \[ C_{total} = .48 + C_{metal2} \text{ pF} \]
   \[ = .48 + (5000 \times 20 \times 10^{-6}) \text{ pF} \]
   \[ = 0.58 \text{ pF} \]

   \[ RC = 250 \times .58 \times 10^{-12} \text{ s} \]
   \[ = .145 \text{ ns} \]

3. We now have 32 registers, so the load capacitance of the registers is

   \[ C_{regfile} = 32 \times C_{reg16} \]
   \[ = 15.36 \text{ pF} \]

   The RC for a 1 μm-wide clock feed is
   \[ = 3.84 \text{ ns} \]

   Hence the clock line has to be widened by 3.84/0.5 or 7.68. For safety one might choose a 10 μm wire.

Now

   \[ C_{total} = 15.36 + C_{metal2} \text{ pF} \]
   \[ = 15.36 + (5000 \times 10 \times 20 \times 10^{-6}) \text{ pF} \]
   \[ = 16.36 \text{ pF} \]

   \[ RC = 25 \times 16.36 \times 10^{-12} \text{ s} \]
   \[ = 0.41 \text{ ns} \]
4.3.8 Wire-length Design Guide

- For sufficiently small wire length, the wire's RC delay can be ignored (i.e., the delay element $\tau$ in Figure 4.17 can be removed) and thus the wire can be modeled as a simple capacitive load.

- To model a wire as a simple capacitive load, the wire’s RC delay $\tau_w$ and gate delay $\tau_g$ must satisfy:

$$\tau_w \ll \tau_g \Rightarrow \frac{1}{2} r c l^2 \ll \tau_g \Rightarrow l \ll \sqrt{\frac{2c}{rg}}$$

- For example, assume $\tau_g = 200\text{ps}$ and for a minimum-width aluminum wire

$$l \ll \sqrt{\frac{2 \times 0.2 \times 10^{-9}}{0.05 / \lambda \times 30 \times 10^{-18} / \lambda}} = 16000\lambda$$

where $r = 0.05 \Omega / \lambda$

$$c = 30 \times 10^{-18} F / \lambda$$

$\lambda$ = is the value of $\lambda$ - design rule

---

**FIGURE 4.17** Simple model for RC delay calculation
Table 4.7 shows the maximum interconnect length for a typical CMOS process in terms of $\lambda$ such that a wire can be modeled as a simple capacitive load. This table assumes gate delays of the order of 100ps to 500 ps.

<table>
<thead>
<tr>
<th>LAYER</th>
<th>MAXIMUM LENGTH ($\lambda$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metal3</td>
<td>10000</td>
</tr>
<tr>
<td>Metal2</td>
<td>8000</td>
</tr>
<tr>
<td>Metal1</td>
<td>5000</td>
</tr>
<tr>
<td>Silicide</td>
<td>600</td>
</tr>
<tr>
<td>Polysilicon</td>
<td>200</td>
</tr>
<tr>
<td>Diffusion</td>
<td>60</td>
</tr>
</tbody>
</table>
• 4.4 Inductance

  - On-chip inductances are small, but bond-wire inductances are large enough to cause troubles for I/O circuits (voltage spike \( dV = L \frac{dI}{dt} \)).

  - For an on-chip conductor whose inductance is
    \[
    L = \frac{\mu}{2\pi} \ln \left( \frac{8h}{w} + \frac{w}{4h} \right)
    \]
    where
    \( w \) = the width of the conductor
    \( h \) = the height above the substrate.

  - For bond-wire inductance
    \[
    L = \frac{\mu}{2\pi} \ln \left( \frac{4h}{d} \right)
    \]
    where
    \( \mu \) = permeability of the wire
    \( h \) = the height above the ground plane
    \( d \) = the diameter of the wire.
4.5 Switching Characteristics

- Referring to Figure 4.18, some timing parameters are defined as follows:
  
  - *Rise time, $t_r$* = time for a waveform to rise from 10% to 90% of its steady-state value.
  - *Fall time, $t_f$* = time for a waveform to fall from 90% to 10% of its steady-state value.
  - *Delay time, $t_d$* = time difference between input transition (50%) and the 50% output level.
  - *Output rising delay time, $t_{dr}$* = the delay time of a rising output in response to the input change.
  - *Output falling delay time, $t_{df}$* = the delay time of a falling output in response to the input change.
4.5.1 Analytic Delay Models

4.5.1.1 Fall Time

- Referring to Figure 4.18, initially the n-device is cut-off and the load capacitor $C_L$ is charged to $V_{DD}$ (at point $X_1$ on the characteristic curve). Application of a step voltage (i.e. $V_{gs} = V_{DD}$) at the input of the inverter changes the operating point to $X_2$. From this onward, the operating point moves toward $X_3$. Thus, it is evident that the fall time $t_f$ consists of two intervals:

1. $t_{f1}$ = period during which the capacitor voltage $V_{out}$ drops from $0.9 \times V_{DD}$ to $(V_{DD} - V_{tn})$. Equivalent circuit is shown in Figure 4.19(a).

2. $t_{f2}$ = period during which the capacitor voltage $V_{out}$ drops from $(V_{DD} - V_{tn})$ to $0.1 \times V_{DD}$.

![Equation](image-url)
During $t_{f1}$, the $n$-device is in saturation, thus

$$C_L \frac{dV_{out}}{dt} = -\frac{\beta_n}{2} (V_{DD} - V_{in})^2$$

solving this equation for $t$, we obtain

$$t_{f1} = \frac{2C_L (V_{in} - 0.1V_{DD})}{\beta_n (V_{DD} - V_{in})^2}$$

During $t_{f2}$, the $n$-device is non-saturated, thus

$$C_L \frac{dV_{out}}{dt} = -\beta_n ((V_{DD} - V_{in})V_{out} - \frac{1}{2}V_{out}^2)$$

Solving this equation, we obtain

$$t_{f2} = \frac{C_L}{\beta_n V_{DD} (1 - n)} \ln(19 - 20n), \text{ where } n = \frac{V_{in}}{V_{DD}}$$

$$t_f = t_{f1} + t_{f2} = \frac{2C_L}{\beta_n V_{DD} (1 - n)} \left[ \frac{(n - 0.1)}{1 - n} + \frac{1}{2} \ln(19 - 20n) \right]$$

(4.37)

that is, $t_f \approx k \frac{C_L}{\beta_n V_{DD}}$ where $k = 3-4$.

- $t_f \propto C_L$, $t_f \propto 1/V_{DD}$ and $t_f \propto \frac{1}{\beta_n}$. 


• 4.5.1.2 Rise Time

- Due to the symmetry of the CMOS circuit, 
  \[
  t_r = \frac{2C_L}{\beta_p V_{DD} (1 - p)} \left[ \frac{p - 0.1}{1 - p} + \frac{1}{2} \ln(19 - 20p) \right] 
  \tag{4.39}
  \]

  where \( p = \frac{|V_{oL}|}{V_{DD}} \). As before, \( t_r \equiv 3 \rightarrow 4 \frac{C_L}{\beta_p V_{DD}} \).

- For equally sized \( n \)- and \( p \)-transistors, where \( \beta_n = 2 \beta_p \), \( t_f = \frac{t_r}{2} \)

- For equal rise and fall time, \( \frac{\beta_n}{\beta_p} = 1 \) This implies \( w_p = 2 \sim 3w_n \), where \( w_p \) is the channel width of the \( p \)-device and \( w_n \) is the channel width of the \( n \)-device.
• 4.5.1.3 Delay Time

- In most CMOS circuits, the delay of a single gate is dominated by the output rise and full times. It is approximated by \( t_{dr} = \frac{t_r}{2} \) and \( t_{df} = \frac{t_f}{2} \).

- The average gate delay for rising and falling transition is \( t_{av} = \frac{t_{df} + t_{dr}}{2} = \frac{t_r + t_f}{4} \).

- Figure 4.20 illustrates a SPICE simulation of a step input applied to an inverter driving a capacitive load. With \( V_{in} = .767V \), \( V_{wp} = -.938V \), \( \beta_n = 4.04 \times 10^{-4} \), \( \beta_p = 3.48 \times 10^{-4} \), \( V_{DD} = 5.0V \), \( C_L = 0.5pF \).
  
  - By Eq.(4.39), \( t_r = 1.04ns \) (compared to \( 1.14ns \) from SPICE)
  
  - By Eq.(4.37), \( t_f = 0.83ns \) (compared to \( 0.89ns \) from SPICE)
  
  - \( t_{dr} = \frac{t_r}{2} = 0.502ns \) (compared to \( 0.52ns \) by SPICE)
  
  - \( t_{df} = \frac{t_f}{2} = 0.41ns \) (compared to \( 0.45ns \) by SPICE).
Alternatively, the output fall delay time can be approximated by \( t_{df} = A_N \frac{C_L}{\beta_n} \) (i.e., \( R_n = \frac{A_N}{\beta_n} \)). \( A_N \) is a process-specific constant and derived as

\[
A_N = \frac{1}{V_{DD}(1-n)} \left( \frac{2n}{1-n} + \ln\left( \frac{2(1-n)-V_O}{V_O} \right) \right)
\]

where \( n = \frac{V_{in}}{V_{DD}} \) and \( V_O = \frac{V_{out}}{V_{DD}} \).

Similarly, \( t_{dr} = A_p \frac{C_L}{\beta_p} \) (E.Q. 4.47) and \( A_p = \frac{1}{V_{DD}(1+p)} \left[ -\frac{2p}{1+p} + \ln\left( \frac{2(1+p)-V_O}{V_O} \right) \right] \) where \( p = \frac{V_{in}}{V_{DD}} < 0 \).

### 4.5.2 Empirical Delay Models

A circuit simulator is employed to find the timing parameter of interest and the coefficient of some delay equation is derived. For example, the coefficients \( A_p \) and \( A_N \) can be found by

\[
A_p = \frac{\beta_p}{C_L} = 0.52 \times 10^{-9} \times \frac{3.48 \times 10^{-4}}{0.5 \times 10^{-12}} = 0.36 \quad (0.31 \text{ calc})
\]

\[
A_N = \frac{\beta_n}{C_L} = 0.45 \times 10^{-9} \times \frac{4.04 \times 10^{-4}}{0.5 \times 10^{-12}} = 0.36 \quad (0.29 \text{ calc})
\]

Note that the values of \( t_{dr-spice} \) and \( t_{df-spice} \) are obtained from SPICE simulation.

Thus, \( t_{dr} = 0.36 \frac{C_L}{\beta_p} \) and \( t_{df} = 0.36 \frac{C_L}{\beta_n} \) for the gates with \( W_p = 2W_n \).
• 4.5.3 Gate delays
  - The delay of simple gate may be approximated by constructing an “equivalent” inverter. The pull-down $n$-transistor and pull-up $p$-transistor of the equivalent inverter are of a size to reflect the effective strength of the real pull-down or pull-up in the gate.

  - For example, the equivalent inverter for the 3-input NAND gate shown in Figure 4.21 has the effective $\beta$ of the $n$-transistor as $\beta_{\text{neff}} = \frac{1}{\frac{1}{\beta_{n1}} + \frac{1}{\beta_{n2}} + \frac{1}{\beta_{n3}}}$

  - For $\beta_{n1} = \beta_{n2} = \beta_{n3}$, $\beta_{\text{neff}} = \frac{\beta_n}{3}$

![Figure 4.21](image-url)
For the pull-up case, only one p-transistor has to turn on to raise the output. Thus, $\beta_{\text{eff}} = \beta_p$.

For $\beta_p = 0.3 \beta_n$, $t_r = k \frac{C_l}{0.3 \beta_n V_{DD}}$, $t_f = k \frac{C_l}{\beta_n V_{DD}}$. Thus $\frac{t_r}{t_f} \approx 1$.

More clearly, for a series of 3 n-transistor, $\beta_{\text{series}} = \frac{\beta_n}{3}$.

In general, the fall time is $mt_f$ for $m$ n-transistors in series; the rise time for $m$ p-transistor in series is $mt_r$.

The fall (rise) time for a parallel connection of $n(p)$-transistors is $t/m (t/m)$ for $m$ transistors in parallel, if all the transistors are turned on simultaneously.
4.5.4 Further Delay Topics

4.5.4.1 Input Waveform Slope

- The analytical delay expression in an inverter presented in the previous section was derived based on the assumption that the input waveform is a step function. However, the real input waveform has a finite slope that will modify the delay of a gate.

- Figure 4.23 shows the influence of input waveform on the delay time of an inverter based on SPICE simulation. The results are tabulated in Table 4.8.

\[ \text{Figure 4.23 Effect of input rise and fall time on inverter delays} \]
The following modification to the output rising delay is made to consider the input waveform slope:

\[ t_{dr} = t_{dr-step} + \frac{t_{input-fall}}{6} (1 - 2p) , \quad p = \frac{V_p}{V_{DD}} < 0 \]

where \( t_{dr-step} \) = the output rising delay due to a step input calculated in EQ. 4.47.

\[ t_{input-fall} \] = the input fall time

Similarly, \( t_{df} = t_{df-step} + \frac{t_{input-rise}}{6} (1 + 2n) \) for \( n = \frac{V_n}{V_{DD}} > 0 \).

These two equations are valid only when \( \frac{t_{input-fall} \beta_n V_{DD}}{C_L} < \frac{6|p|}{(1 - |p|)^3} \) and \( \frac{t_{input-fall} \beta_n V_{DD}}{C_L} < \frac{6n}{(1 - n)^3} \).
• **4.5.4.2 Input Capacitance**
  
  – The input capacitance is in fact a function of gate voltage. (i.e., it is not a constant)
  
  – An effect known as bootstrapping can also modify the effective input capacitance of a logic gate.
  
  – As shown in Figure 4.24(a), in the case where the input is rising (that is, the output is high), the effective input capacitance is $C_{gs} + C_{gd}$. When the output starts to fall, the voltage across $C_{gd}$ increases, requiring the input to supply more current to charge $C_{gd}$. This effect is seen in Figure 4.24(b).

![Figure 4.24](image-url)
• **4.5.4.3 Switch Level Models**
  
  – A switch model represents a transistor as a resistance for charging or discharging a capacitance is shown in Figure 4.25.

  ![Figure 4.25](image)

  **FIGURE 4.25**
  A switch-level RC model

  – A variety of timing models have developed to estimate the delay of logic gates:

    • *Simple RC model*: the total resistance of pull-up or pull-down path is calculated and all the capacitance of nodes are lumped onto the output of the gate.

    For the example shown in Figure 4.26.,

    \[
    t_{df} = \sum R_{\text{pull-down}} \times \sum C_{\text{pull-down-path}} = (R_{N1} + R_{N2} + R_{N3} + R_{N4}) \times (C_{\text{out}} + C_{ab} + C_{bc} + C_{cd})
    \]

    , while for the rise delay \( t_{dr} = R_{p4} \times C_{\text{out}} \). This is a pessimistic model.
• **Penfield-Rubenstein Model**: was developed to calculate delays in generalized RC trees. \( t_d = \sum_i R_i C_i \) for a group of series transistors. Where \( R_i \) is the summed resistance from point \( i \) to power or ground and \( C_i \) is the capacitance at point \( i \).

For example, for the gate shown in Figure 4.26,

\[
t_{df} = R_{N_1} * C_{cd} + (R_{N_1} + R_{N_2}) * C_{bc} + (R_{N_1} + R_{N_2} + R_{N_3}) * C_{ab} + (R_{N_1} + R_{N_2} + R_{N_3} + R_{N_4}) * C_{out}
\]

• **Penfield-Rubenstein Slope Model**: widely used in transistor-level-timing analyzers. See the textbook for details.
• 4.5.4.4 Macromodeling
  
  To derive a set of accurate formulae to calculate gate capacitance and logic gate behavior based on the device equation.

  Figure 4.28 shows a typical model along with the timing model, where $t_{\text{swin}}$ is the input waveform, $t_{\text{swout}}$ is the output waveform, $C_{\text{in}}$ is the input capacitance, and $C_L$ is the output capacitance.

  Another commonly used approach in ASIC community treats logic gates as simple delay elements. Each gate type is simulated with a circuit simulator to derive an equation to compute the delay of a particular gate: $t_d = t_{\text{internal}} + k \times t_{\text{output}}$, where $t_{\text{internal}}$ is a fixed delay when no load is attached; $k$ is the output loading; $t_{\text{output}}$ is the output delay per output loading.

![Diagram](image.png)  

**FIGURE 4.28** Model used in macromodeling approach
- Figure 4.29 shows a typical SPICE circuit used to calibrate delay equation.
- Table 4.9 summarizes the data.
- Thus for the gate shown in Figure 4.29, \( t_{dr} = (0.255 + k \times 2.12) \) ns, \( t_{df} = (0.42 + k \times 3.82) \) ns

![FIGURE 4.29 SPICE circuit and results for delay modeling on a 3-input NAND gate](image)

<table>
<thead>
<tr>
<th>TABLE 4.9 NAND3 SPICE Delays</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOAD CONDITIONS</td>
</tr>
<tr>
<td>Time</td>
</tr>
<tr>
<td>( t_{dr} ) (ns)</td>
</tr>
<tr>
<td>( t_{df} ) (ns)</td>
</tr>
<tr>
<td>( t_{output _rise} ) (ns/pF)</td>
</tr>
<tr>
<td>( t_{output _fall} ) (ns/pF)</td>
</tr>
</tbody>
</table>
4.5.4.5 Body Effect

Body effect is the term given to the modification of the threshold voltage, $V_t$, with a voltage difference between source and substrate.

$$\Delta V_t \propto r \sqrt{V_{sb}}$$

where $r$ is a constant, $V_{sb}$ is the voltage between source and substrate, and $\Delta V_t$ is the change in threshold voltage.

For the example shown in Figure 4.30(a), the n-transistor at the output will switch slower if the source potential of this transistor is not the same as the substrate.

In the upper NAND gate the lower transistors are initially turned on while transistor $N_{4A}$ is turned off. The result is seen in Figure 4.30(b) in the form of waveform $CD$ when the input on $N_{4A}$ rises.

In the lower NAND gate, the upper transistors are turned on initially, while transistor $N_{1B}$ is turned off. (Effectively, the $V_{sb}$ of the upper transistors is not equal to zero). The result is seen in Figure 4.30(b) when $N_{1B}$ turns on, which indicates the output transition lags behind $0.4 \text{ ns}$.

Two strategies to offset the body effect:

- Place the transistors with the latest arriving signals nearest the output of a gate.
- Minimize the capacitance of internal nodes.
FIGURE 4.30 SPICE circuit for observing the result of body effect on gate delay.
• 4.6 CMOS-Gate Transistor Sizing

• 4.6.1 Cascaded Complementary Inverters
  – Previously discussed, we must take $W_p = (2 \rightarrow 3) W_n$ to approximately equalize the rise and fall times of an inverter. However, in some cascaded structures it is possible to use minimum or equal-sized devices to achieve the same result.
  
  – For example, as for the circuit shown in Figure 4.31(a), with $W_p = 2W_n$,
  
  \[
  t_{\text{inv-pair}} \propto t_{\text{fall}} + t_{\text{rise}} \propto R3C_{eq} + \frac{2R}{2}3C_{eq} = 3RC_{eq} + 3RC_{eq} = 6RC_{eq},
  \]

  where $R$ is the effective “ON” resistance of a unit-sized $n$-transistor ($W=2, L=1$), and $C_{eq} = C_g + C_d$ is the capacitance of a unit-sized gate and drain region. As for the case shown in Figure 4.31(b) with $W_p = W_n$,

  \[
  t_{\text{inv-pair}} \propto t_{\text{fall}} + t_{\text{rise}} \propto R2C_{eq} + 2R2C_{eq} = 6RC_{eq}
  \]

![Diagram of CMOS inverter pair timing response](image)
Also note that changes in the $\beta$ ratio affect the inverter logic threshold voltage $V_{\text{inv}}$, which directly influences the delay of output response. From equation (2.2),

$$V_{\text{inv}} = \frac{V_{\text{DD}} + V_{\text{tp}} + V_{\text{tn}} \sqrt{\frac{\beta_n}{\beta_p}}}{1 + \sqrt{\frac{\beta_n}{\beta_p}}}$$

Table 4.10 summarizes $V_{\text{inv}}$ for a range of its constituent parameters.

<table>
<thead>
<tr>
<th>$V_{\text{DD}}$</th>
<th>$V_{\text{tn}}$</th>
<th>$V_{\text{tp}}$</th>
<th>$\beta_n$</th>
<th>$\beta_p$</th>
<th>$V_{\text{inv}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>.7</td>
<td>-0.7</td>
<td>1</td>
<td>1</td>
<td>2.5</td>
</tr>
<tr>
<td>5</td>
<td>.7</td>
<td>-0.7</td>
<td>.5</td>
<td>1</td>
<td>2.8</td>
</tr>
<tr>
<td>5</td>
<td>.7</td>
<td>-0.7</td>
<td>1</td>
<td>.5</td>
<td>2.2</td>
</tr>
<tr>
<td>3</td>
<td>.5</td>
<td>-0.5</td>
<td>1</td>
<td>1</td>
<td>1.5</td>
</tr>
<tr>
<td>3</td>
<td>.5</td>
<td>-0.5</td>
<td>.5</td>
<td>1</td>
<td>1.67</td>
</tr>
<tr>
<td>3</td>
<td>.5</td>
<td>-0.5</td>
<td>1</td>
<td>.5</td>
<td>1.32</td>
</tr>
</tbody>
</table>
4.6.3 Stage Ratio

A chain of increasingly larger inverters is usually employed to drive large loads. The ratio by which each stage is increased in size is called the stage ratio.

Considering the circuit shown in Figure 4.33, \( \text{inv-1} \) is a minimum-sized inverter driving \( \text{inv-2} \), which is \( a \) times the size of \( \text{inv-1} \). Similarly, the size of \( \text{inv-3} \) is \( a \) times the size of \( \text{inv-2} \) (that is, \( \text{inv-3} \) is \( a^2 \) the size of \( \text{inv-1} \)).

The delay through each stage is \( at_d \), where \( t_d \) is the average delay of a minimum-sized inverter driving another minimum-sized inverter. Hence the delay through \( n \) stages is \( nat_d \). If the ratio of the load capacitance to the capacitance of a minimum-sized inverter is \( R=C_L/C_g \), then \( R = a^n \). Hence \( \ln(R) = n \ln(a) \).

Thus, \( \text{Total delay} = nat_d = \frac{\ln(R)}{\ln(a)} at_d \).

The variable part of the above equation, normalized to \( e \), is graphed in Figure 4.33(b). The graph shows when \( a = 2.7 = e \) would minimize the total delay.
FIGURE 4.33  Stage ratio
(a) circuit; (b) graph
4.7 Power Dissipation

- Two components of CMOS power dissipation:
  - Static dissipation due to leakage current.
  - Dynamic dissipation due to switching transient current and charging/discharging of load capacitance.

4.7.1 Static Dissipation

- For the inverter shown in Figure 3.34, if the input is either at one of its steady-state logic values (0 or 1), there is no direct path from V_{DD} to V_{SS}. Thus no static power \( P_s \) dissipates, theoretically.

- However, some small static dissipation actually exists due to:
  - Reverse biased leakage between diffusion regions and the substrate.
  - Subthreshold conduction (i.e. \( V_{in} < |V_i| \))

\[
\begin{align*}
V_{in} &= 0 & V_{out} &= 1 \\
V_{in} &= 1 & V_{out} &= 0
\end{align*}
\]

**FIGURE 4.34**
CMOS inverter model for static power dissipation evaluation.
The leakage current between diffusion and substrate can be explained by the model shown in Figure 3.35. The diodes in the model are reverse-biased and the leakage current is described by diode equation.

\[ i_o = i_s (e^{qV/kT} - 1) \]

Where  
- \( i_s \) = reverse saturation current  
- \( V \) = diode voltage  
- \( q \) = electronic charge (\( 1.602 \times 10^{-19} \) C)  
- \( K \) = Boltzmann’s constant (\( 1.38 \times 10^{-23} \) J/K)  
- \( T \) = temperature

**FIGURE 4.35** Model describing parasitic diodes present in a CMOS inverter.
– The leakage current per device ranges from 0.1\text{nA} to 0.5\text{nA}.
– Total static power dissipation $P_s$ can be obtained by

$$P_s = \sum_{i=1}^{n} \text{leakage current} \times \text{supply voltage}$$

where $n =$ number of devices.

**Example**

For a process with $\beta_p$ of 30 \text{\mu A}/\text{V}^2 and a $\beta_n$ of 85 \text{\mu A}/\text{V}^2 ($V_{th} = |V_{tp}| = 0.7\text{V}, V_{DD} = 5\text{V}$), calculate the static power dissipation of a $32 \times 32$ ROM which contains a 1:32 pseudo-nMOS row decoder and pMOS pull-ups on the 32-bit lines. The aspect ratio of all pMOS pull-ups (W/L) is 1. Each pMOS load can source $(\beta(V_{gs} - V_t)^2)/2$ of current.

$$I_{load} = \left(30 \times (5 - 0.7)^2\right)/2 \mu\text{A} = 277\mu\text{A}$$

$$P_{load} = 1.4 \text{ mW} = (277\mu\text{A} \times 5\text{V})$$

Assuming that one row decoder is on and 50\% of the bit lines are on at any one time yields

$$P_{total} = 17 \times 1.4 \text{ mW} = 23.6 \text{ mW}.$$
• 4.7.2 Dynamic Dissipation
  
  - Consists of short circuit power and load capacitance charging/discharging power.
  
  - Short circuit dissipation is due to existence of a direct path from $V_{DD}$ to $V_{SS}$ when the output changes either from 1 to 0 or 0 to 1.
  
  - Short circuit dissipation depends on the input rise/fall time, the load capacitance and gate design.
  
  - Fig 4.36 depicts a scenario about how output loading could influence the short circuit current.
  
  - The average dynamic power $P_d$ dissipated during switching for a square-wave input, $V_{in}$, having a repetition frequency of $f_p = 1/t_p$ is given by

$$P_d = \frac{1}{t_p} \int_{t/2}^{t_p} i_n(t) V_{out} dt + \frac{1}{t_p} \int_{t/2}^{t_p} i_p(t)(V_{DD} - V_{out}) dt$$

For a step input and with $i_n(t) = \frac{C_L dV_{out}}{dt}$,

$$P_d = \frac{C_L}{t_p} \int_0^{V_{DD}} V_{out} dV_{out} + \frac{C_L}{t_p} \int_0^{V_{DD}} (V_{DD} - V_{out}) d(V_{DD} - V_{out}) = \frac{C_L V_{DD}^2}{t_p} = C_L V_{DD}^2 f_p$$

  - Dynamic power dissipation can be limited by reducing supply voltage, output capacitance and the switching frequency.
FIGURE 4.36 SPICE circuits and results showing dynamic short-circuit current and capacitive current for a CMOS inverter for varying load capacitances (the 0V voltage sources are used to measure currents)
• **4.7.3 Short-circuit Dissipation**
  
  For an inverter without load, assuming $t_r = t_f = t_{rf}$, the short circuit power dissipation (detailed derivation can be found in Weste textbook).
  
  $$P_{sc} = \frac{\beta}{12}(V_{DD} - 2V_t)^3 \frac{t_{rf}}{t_p}$$

• **4.7.4 Total Power Dissipation**
  
  $P_{total} = P_s + P_d + P_{sc}$

  More realistic situation, the dynamic power $P_d$ should be calculated as follows:

  $$P_d = \frac{Percentage\text{-}activity \times C_{Total} \times V_{DD}^2}{t_p},$$

  where *percentage-activity* is a ratio between the estimated number of switchings and the number of clock cycles during a certain period of time.
4.8 Sizing Routing Conductors

- Metal power-carrying conductors have to be sized for three reasons:
  - *Metal migration* (electro-migration)
  - *Power supply noise and integrity* (i.e., satisfactory power and signal voltage levels are presented to each gate.)
  - *RC delay*

- Metal migration is the transport of metal ions through a conductor. It may cause breakage or deformation of conductor.

- Factors that influence the electro-migration rate are:
  - *Current density*, as a rule of thumb, 0.4 mA/um to 0.1 mA/um of metal width should be used for both $V_{DD}$ and $V_{SS}$ lines.
  - *Temperature*
  - *Crystal structure*

- Existence of substantial resistance on supply or ground line would cause considerable $IR$ drop (power drop and ground bounce) during charging transients. This directly reduces the noise margin of gates and causes incorrect operation of gates.
4.8.1 Power Drop and Ground Bounce

For a clocked synchronous system, the output transitions of the gates are made close to the clock transition. Thus, a large current spike would appear on power or ground bus. This would effectively reduce the noise margin. (see the following figure)
Example

What would be the conductor width of power and ground wires to a 50 MHz clock buffer that drives 100pF of on-chip load to satisfy the metal-migration consideration ($J_{AL} = 0.5$ mA/µ)? What is the ground bounce with the chosen conductor size? The module is 500µ from both the power and ground pads and the supply voltage is 5 volts. The rise/fall time of the clock is 1 ns.

1. \[ P = CV_{DD}^2f \]
\[ = 100 \times 10^{-12} \times 25 \times 50 \times 10^6 \]
\[ = 125 \text{ mW} \]

\[ I = 25 \text{ mA} \]

Thus the width of the clock wires should be at least 60µ. A good choice would be 100µ.

2. \[ R = \frac{500}{100 \times .05} \]
\[ = 5 \text{ squares} \times .05 \text{ } \Omega/\text{sq.} \]
\[ = .25 \Omega \]

\[ IR = \frac{C_dV}{dt}R = \frac{100 \times 10^{-12} \times 5}{1 \times 10^{-9}} \times .25 \]
\[ = 125 \text{ mV} \text{ (also see Section 5.5.16)} \]
• 4.8.2 Contact Replication
  – For a very wide conductor to connect to another layer, an array of small contacts, suitably spaced, generally provides just as much current-carrying capacity as a single long, narrow contact. (see Figure 4.38)

  ![Contact Replication Diagram](image)

  **FIGURE 4.38** Contact structures for linear and orthogonal joints

• 4.9 Charge Sharing
  – In many structures a bus can be modeled as a capacitor $C_b$ as shown in Figure 4.39. Sometimes the voltage on this bus is sampled (latched) to determine the state of a given signal. Charge sharing thus occurs.

  ![Charge Sharing Diagram](image)

  **FIGURE 4.39** Charge-sharing mechanism
For a successful sampling, the resultant voltage $V_R$ (not shown) should be correctly reflect the state of the bus. For example, $Q_b = C_b V_b$, $Q_s = C_s V_s$, $Q_T = C_b V_b + C_s V_s$, $C_T = C_b + C_s$. Thus $V_R = Q_T / C_T = (C_b V_b + C_s V_s) / (C_b + C_s)$, if $V_b = V_{DD}$ (the bus is in state ‘1’), and $V_b >> V_s$, then $V_R = V_{DD} [C_b / (C_b + C_s)]$. For a successful sampling, $C_s << C_b$. This results in $V_R \approx V_{DD}$.

Example: A precharge bus has a loading of $10 pF$. At a point in the clock cycle, 64 registers with transmission gates on their inputs turn on. The input load of each register (after the transmission gate) is $1 pF$. Calculate the change in precharge voltage.

What would be an alternative approach?

1. Here $C_b = 10 pF$

   $C_s = 64 \times 1 pF = 6.4 pF$

   $V_{DD} = 5 V$

   Hence

   $$V_R = 5 \times \frac{10}{10 + 6.4}$$

   $$= 3.05 \text{ volts (change in voltage is 1.9V)}.$$

2. The most obvious approach to alleviating the problem is to use buffer inverters on the input of each register. (The above example would probably point to a very suspect design approach!)
• 4.10 Design Margining
  – The circuit behavior can be affected by the variation of operating conditions and manufacturing process.
    These variations include:
    • *operating temperature variation*
    • *supply voltage variation*
    • *process variation*

• 4.10.1 Temperature Variation
  – As the temperature is increased, the drain current is reduced. This variation is shown in Figure 4.40.

![Figure 4.40](image)

- The temperature variation requirement for different applications is shown as follows:

<table>
<thead>
<tr>
<th>Application</th>
<th>Temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Min (°C)</td>
</tr>
<tr>
<td>Commercial</td>
<td>0</td>
</tr>
<tr>
<td>Industrial</td>
<td>-40</td>
</tr>
<tr>
<td>Military</td>
<td>-55</td>
</tr>
</tbody>
</table>
The die temperature is specified as follows: \( T_j = T_a + Q_{ja} \times P_d \)

where
- \( T_j \) = the junction temperature in °C (temperature of the chip itself)
- \( T_a \) = the ambient temperature in °C (temperature of surrounding air)
- \( Q_{ja} \) = the package thermal impedance, expressed in °C/watt
- \( P_d \) = the power dissipation

For example, if \( Q_{ja} = 30 \, ^\circ C/watt \), \( P_d = 1 \, watt \), and \( T_a = 85 \, ^\circ C \), then \( T_j = 115 \, ^\circ C \).

- **4.10.2 Supply Voltage Variation**
  - When specifying a part, a variation of 10% (normally) on the supply voltage accompanies the data sheet.

- **4.10.3 Process Variation**
  - The variation in device characteristics due to process variation follows a normal distribution as shown in Figure 4.41.
The variations in device performance can be caused by variations in doping density, implant dose, and variations in the width and thickness of active diffusion and oxide layers and passive conductors.

When considering the influence of process variation on the transistor speed, the following terms are used:

- nominal (typical)
- fast
- slow

There are four types of boundary conditions for two types of transistors:

- Fast-n and fast-p
- Fast-n and slow-p
- Slow-n and fast-p
- Slow-n and slow-p

If the gains of the p- and n-transistors track but the threshold voltage is not, the following process corners can be observed:

- Slow-n and low-$V_{tp}$
- Low-$V_{tn}$ and slow-p
• 10.10.4 Design Corners
  
  - A design corner refers to an imaginary box that surrounds the guaranteed performance of the transistors.
  
  - The worst-power or high-speed corner: fast-n and fast-p process corner combined with the lowest operating temperature and the highest operating voltage.
  
  - The worst-speed corner: slow-n and slow-p process corner combined with the highest operating temperature and the lowest operating voltage.
  
  - The worst-speed corner can be used to check external setup times while the highest speed corner can be used to check hold time constraints.
  
  - Table 4.11 shows a list of checks performed for a CMOS digital system.

<table>
<thead>
<tr>
<th>PROCESS</th>
<th>TEMP</th>
<th>VOLTAGE</th>
<th>TESTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fast-n/fast-p</td>
<td>0°C</td>
<td>5.5V (3.6V)</td>
<td>Power dissipation (DC), clock races, hold time constraints</td>
</tr>
<tr>
<td>Slow-n/slow-p</td>
<td>125°C</td>
<td>4.5V (3.0V)</td>
<td>Circuit speed, setup time constraints</td>
</tr>
<tr>
<td>Slow-n/fast-p</td>
<td>0°C</td>
<td>5.5V (3.6V)</td>
<td>Pseudo-nMOS noise margin, level shifters, memory write/read, ratioed circuits</td>
</tr>
<tr>
<td>Fast-n/slow-p</td>
<td>0°C</td>
<td>5.5V (3.6V)</td>
<td>Memories, ratioed circuits, level shifters</td>
</tr>
</tbody>
</table>
• 4.10.5 Packaging Issues
  – Selection of package must consider the thermal impedance of the package and the inductance of package pins.
  – The thermal impedance is a measure of the effectiveness with which a package can conduct heat away from the die.
  – Figure 4.42 shows some typical packages.

**FIGURE 4.42** Typical packages used for CMOS chips
• **4.11 Yield**
  
  The yield is influenced by such factors as:
  
  • technology
  • chip area
  • layout

  - Yield is defined as \( Y = \frac{\text{No. of good chips on wafer}}{\text{Total number of chips}} \)

• **4.12 Reliability**
  
  The potential reliability of a CMOS chip:
  
  • “Hot electron” effects
  • Electro-migration
  • Oxide failure
  • Die temperature
  • ESD protection.
4.13 Scaling of MOS-transistor Dimensions

- **Scaling**: to shrink the size of physical dimensions or to scale the property values of a transistor by a dimension factor $\alpha$.

- Three types of scaling:
  - **Constant field scaling**: involve scaling of all physical dimensions (including those vertical to surface), device voltages and the concentration densities.
  - **Constant voltage scaling**: where $V_{DD}$ voltage is kept constant, while the process is scaled.
  - **Lateral scaling**: where only gate length is scaled.

- The resultant effect of these three types of scaling is shown in Table 4.12.

### Table 4.12 Influence of Scaling on MOS-Device Characteristics

<table>
<thead>
<tr>
<th>PARAMETER</th>
<th>Constant field</th>
<th>Constant voltage</th>
<th>Lateral</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length ($L$)</td>
<td>$1/\alpha$</td>
<td>$1/\alpha$</td>
<td>$1/\alpha$</td>
</tr>
<tr>
<td>Width ($W$)</td>
<td>$1/\alpha$</td>
<td>$1/\alpha$</td>
<td>$1/\alpha$</td>
</tr>
<tr>
<td>Supply voltage ($V$)</td>
<td>$1/\alpha$</td>
<td>$1$</td>
<td>$1/\alpha$</td>
</tr>
<tr>
<td>Gate-oxide thickness ($t_o$)</td>
<td>$1/\alpha$</td>
<td>$1/\alpha$</td>
<td>$1$</td>
</tr>
<tr>
<td>Current ($I = (W/L)(1/\mu_C)V^2$)</td>
<td>$1/\alpha$</td>
<td>$\alpha$</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>Transconductance ($g_{m}$)</td>
<td>$1$</td>
<td>$1/\alpha$</td>
<td>$1$</td>
</tr>
<tr>
<td>Junction depth ($X_J$)</td>
<td>$1/\alpha$</td>
<td>$1/\alpha$</td>
<td>$1$</td>
</tr>
<tr>
<td>Substrate doping ($N_d$)</td>
<td>$\alpha$</td>
<td>$\alpha$</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>Electric Field across gate oxide ($E$)</td>
<td>$1$</td>
<td>$1$</td>
<td>$1$</td>
</tr>
<tr>
<td>Depletion layer thickness ($d$)</td>
<td>$1/\alpha$</td>
<td>$1/\alpha$</td>
<td>$1$</td>
</tr>
<tr>
<td>Load Capacitance ($C = WL/\mu_C$)</td>
<td>$1/\alpha$</td>
<td>$1/\alpha^2$</td>
<td>$1/\alpha$</td>
</tr>
<tr>
<td>Gate Delay ($VC/\mu$)</td>
<td>$1/\alpha$</td>
<td>$1/\alpha^2$</td>
<td>$1/\alpha^2$</td>
</tr>
</tbody>
</table>

**RESULTANT INFLUENCE**

- DC power dissipation ($P_d$) $1/\alpha^2$, $\alpha$, $\alpha$
- Dynamic power dissipation ($P_d$) $1/\alpha^2$, $\alpha$, $\alpha$
- Power-delay product $1/\alpha^3$, $1/\alpha$, $1/\alpha$
- Gate Area ($A = WL$) $1/\alpha^2$, $1/\alpha^2$, $1/\alpha$
- Power Density ($V/d$) $1/\alpha^3$, $\alpha^3$, $\alpha^3$
- Current Density $\alpha$, $\alpha^3$, $\alpha^2$
• 4.13.2 Interconnect-Layer scaling

  - Scaling of interconnect could result in changes of interconnect resistance and capacitance.
  
  - For example, scaling the thickness and width of a conductor by \( \alpha \), the scaled-line resistance \( R' \) is given by

\[
R' = R_{s}[\frac{L/L}{W/W}] = \rho \cdot \frac{L/L}{t/t} = \alpha R,
\]

where \( \rho \) = conductivity

\[
R'_{s} = \frac{\rho}{t/\alpha}
\]

  - The sheet resistance of the scaled-line

\( t = \) conductor thickness before scaling

  - The voltage drop along the scaled-line for a constant field scaling is

\[
V_{d}' = \frac{I}{\alpha} (\alpha R) = IR
\]

  - The line-response time is

\[
t'_{s} = (\alpha R) \frac{C}{\alpha} = RC
\]
– The influence of scaling on interconnect, if the interconnect is scaled by $\alpha$ and the current is increased by $\frac{1}{\alpha}$, is summarized in Table 4.13.

<table>
<thead>
<tr>
<th>PARAMETERS</th>
<th>SCALING FACTOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Line resistance ($r$)</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>Line response ($re$)</td>
<td>1</td>
</tr>
<tr>
<td>Voltage drop</td>
<td>1</td>
</tr>
</tbody>
</table>

### 4.13.3 Scaling in Practice

– As it can be seen from Table 4.14, constant voltage scaling has been used in the past.

<table>
<thead>
<tr>
<th>YEAR</th>
<th>TECH</th>
<th>CHIP</th>
<th>SIZE ($T_1$)</th>
<th>SIZE (mm$^2$)</th>
<th>SPEED (MHz)</th>
<th>$V_{DD}$</th>
<th>TYPE OF SCALING</th>
</tr>
</thead>
<tbody>
<tr>
<td>1980-1984</td>
<td>3.5μ</td>
<td>16-bit datapath and RAM</td>
<td>12K</td>
<td>25</td>
<td>5</td>
<td>5</td>
<td>Constant voltage</td>
</tr>
<tr>
<td>1985</td>
<td>2.0μ</td>
<td>Lisp μProcessor</td>
<td>250K</td>
<td>225</td>
<td>5</td>
<td>5</td>
<td>Constant voltage</td>
</tr>
<tr>
<td>1987</td>
<td>1.5μ</td>
<td>Lisp μProcessor</td>
<td>250K</td>
<td>144</td>
<td>8</td>
<td>5</td>
<td>Constant voltage</td>
</tr>
<tr>
<td>1989</td>
<td>1.2μ</td>
<td>Lisp μProcessor</td>
<td>250K</td>
<td>100</td>
<td>12</td>
<td>5</td>
<td>Constant voltage</td>
</tr>
<tr>
<td>1990</td>
<td>1.0μ</td>
<td>Ghost canceller</td>
<td>500K</td>
<td>54</td>
<td>56</td>
<td>3–5</td>
<td>Constant voltage</td>
</tr>
<tr>
<td>1992</td>
<td>0.8μ</td>
<td>Video decoder</td>
<td>1.2M</td>
<td>120</td>
<td>40</td>
<td>5</td>
<td>Constant voltage</td>
</tr>
<tr>
<td>1993+</td>
<td>0.5μ</td>
<td>??</td>
<td>&gt;1M</td>
<td>100+</td>
<td>100+</td>
<td>3.3</td>
<td>Scaled $V_{DD}$ and gate length</td>
</tr>
</tbody>
</table>
• 4.14 Summary
  – resistance, capacitance, inductance calculations
  – delay estimation
  – power estimation
  – design margining and reliability
  – effect of scaling