CMOS

Digital Integrated Circuits

Analysis and Design

Chapter 6

MOS Inverters: Switching Characteristics and Interconnect Effects

Introduction

- The parasitic capacitance associated with MOSFET
  - C_{gd}, C_{gs}, gate overlap with diffusion
  - C_{db}, C_{sb}, voltage dependant junction capacitance
  - C_g, the thin-oxide capacitance over the gate area
  - C_{int}, the limped interconnect capacitance
  - Load capacitance $C_{load}= C_{gd,n} + C_{gd,p} + C_{db,n} + C_{db,p} + C_{int} + C_g$
    - $C_{sb,n}$ and $C_{sb,p}$ have no effect on the transient behavior of the circuit, since $V_{SB}=0$
    - The delay times calculated using C_{load} may slightly overestimate the actual inverter delay
      - Charging, discharging

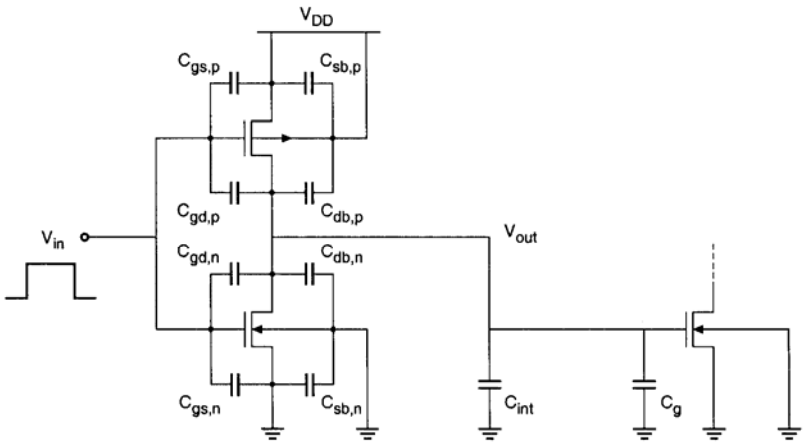


Figure 6.1 Cascaded CMOS inverter stages.

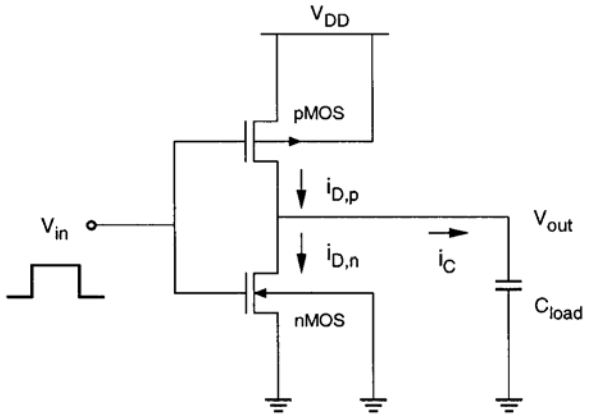


Figure 6.2 First-stage CMOS inverter with lumped output load capacitance.

Delay-time definitions

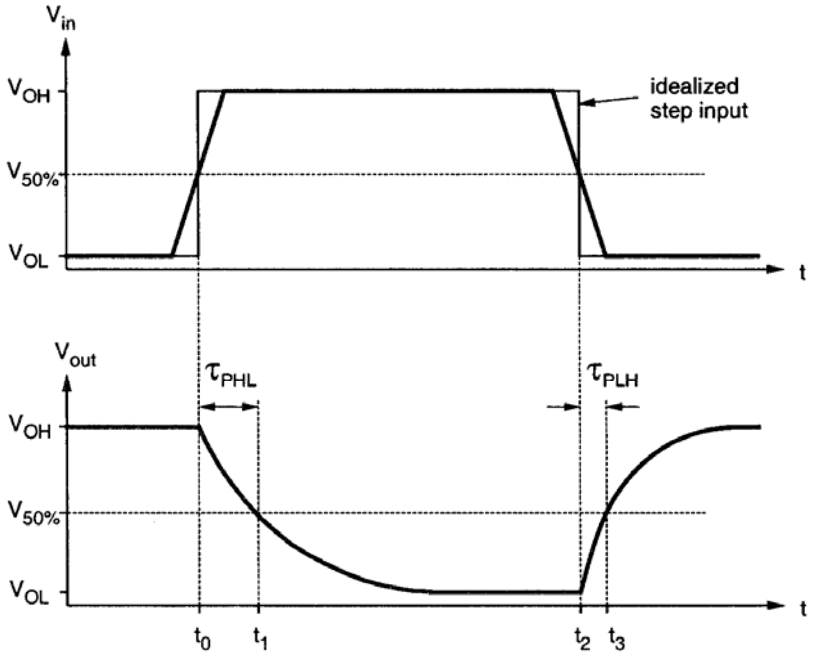


Figure 6.3 Input and output voltage waveforms of a typical inverter, and the definitions of propagation delay times. The input voltage waveform is idealized as a step pulse for simplicity.

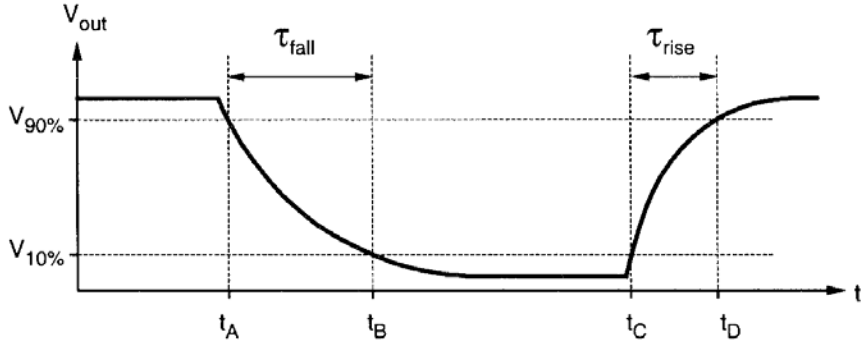


Figure 6.4 Output voltage rise and fall times.

$$V_{50\%} = V_{OL} + 1/2(V_{OH} - V_{OL}) = 1/2(V_{OH} + V_{OL})$$

$$\tau_{PHL} = t_1 - t_0$$

$$\tau_{PLH} = t_3 - t_2$$

$$\tau_P = \frac{\tau_{PHL} + \tau_{PLH}}{2}$$

$$V_{10\%} = V_{OL} + 0.1 \cdot (V_{OH} - V_{OL})$$

$$V_{90\%} = V_{OL} + 0.9 \cdot (V_{OH} - V_{OL})$$

$$\tau_{fall} = t_B - t_A$$

$$\tau_{rise} = t_D - t_C$$

Calculation of delay time

- The simplest approach for calculating the propagation delay times τ_{PHL} and τ_{PLH}
  - Estimating the average capacitance current during charge down and charge up
  - $$\tau_{PHL} = \frac{C_{load} \cdot \Delta V_{HL}}{I_{avg,\,HL}} = \frac{C_{load} \cdot (V_{OH} - V_{50\%})}{I_{avg,\,HL}}$$

    $$\tau_{PLH} = \frac{C_{load} \cdot \Delta V_{LH}}{I_{avg,\,LH}} = \frac{C_{load} \cdot (V_{50\%} - V_{OL})}{I_{avg,\,LH}}$$

    $$I_{avg,\,HL} = \frac{1}{2}[i_C(V_{in} = V_{OH}, V_{out} = V_{OH}) + i_C(V_{in} = V_{OH}, V_{out} = V_{50\%})$$

    $$I_{avg,\,LH} = \frac{1}{2}[i_C(V_{in} = V_{OL}, V_{out} = V_{50\%}) + i_C(V_{in} = V_{OL}, V_{out} = V_{OL})]$$
  - Not very accurate estimate of the delay time

Calculation of delay time(1)

- The propagation delay times can be found more accurately by solving the state equation of the output node in the time domain

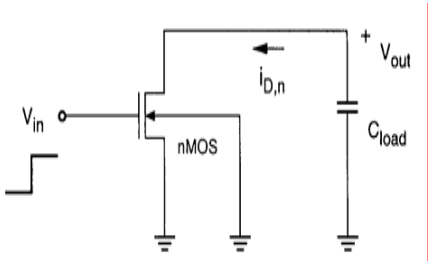


Figure 6.5 Equivalent circuit of the CMOS inverter during high-to-low output transition.

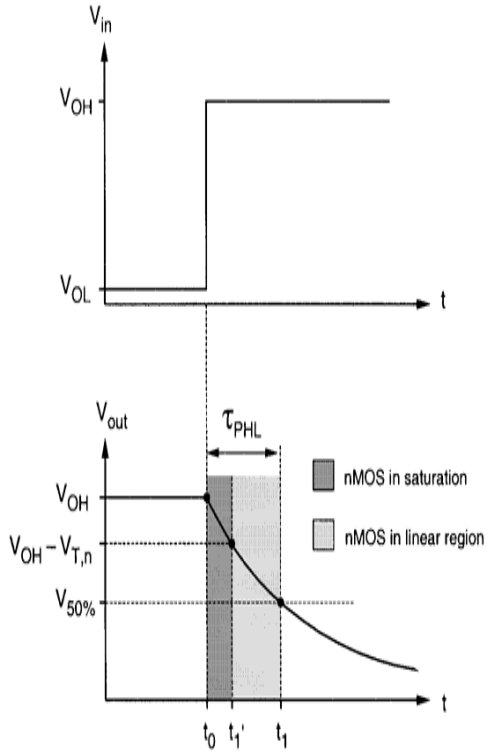


Figure 6.6 Input and output voltage waveforms during high-to-low transition.

$$C_{load}\frac{dV_{out}}{dt} = i_C = i_{D,p} - i_{D,n}$$

First, we consider the resing-input case for a CMOS inverter

the nMOS on $\Rightarrow$ starting to discharge, pMOS off

$$i_{D,p} \approx 0,\ C_{load}\frac{dV_{out}}{dt} = -i_{D,n}$$

$$i_{D,n} = \frac{k_n}{2}\left(V_{in} - V_{T,n}\right)^2 = \frac{k_n}{2}\left(V_{OH} - V_{T,n}\right)^2 \text{ for } V_{OH}\text{-}V_{T,n} < V_{out} \le V_{OH}$$

$$\int_{t=t_0}^{t=t_1'} dt = -C_{load}\int_{V_{out}=V_{OH}}^{V_{out}=V_{OH}-V_{T,n}}\left(\frac{1}{i_{D,n}}\right)dV_{out} = -\frac{2C_{load}}{k_n\left(V_{OH}-V_{T,n}\right)^2}\int_{V_{out}=V_{OH}}^{V_{out}=V_{OH}-V_{T,n}} dV_{out}$$

$$\Rightarrow t_1' - t_0 = \frac{2C_{load}V_{T,n}}{k_n\left(V_{OH}-V_{T,n}\right)^2}$$

At $t = t_1'$, the output voltage is $(V_{DD}\text{-}V_{T,n})$ and the transistor will be at the saturation-linear region boundary. Considering nMOS linear

$$i_{D,n} = \frac{k_n}{2}\left[2\left(V_{in}-V_{T,n}\right)V_{out} - V_{out}^2\right] = \frac{k_n}{2}\left[2\left(V_{OH}-V_{T,n}\right)V_{out} - V_{out}^2\right] for V_{out} \le V_{OH} - V_{T,n}$$

$$\int_{t=t_{1'}}^{t=t_1} dt = -C_{load}\int_{V_{out}=V_{OH}-V_{T,n}}^{V_{out}=V_{50\%}}\left(\frac{1}{i_{D,n}}\right)dV_{out} = -2C_{load}\int_{V_{out}=V_{OH}-V_{T,n}}^{V_{out}=V_{50\%}}\left(\frac{1}{k_n[2(VOH-V_{T,n})V_{out}-V_{out}^2]}\right)dV_{ou}$$

$$t_1 - t_{1'} = -\frac{2C_{load}}{k_n}\frac{1}{2(V_{OH}-V_{T,n})}\ln\left(\frac{V_{out}}{2(V_{OH}-V_{T,n})-V_{out}}\right)$$

$$t_1 - t_{1'} = \frac{C_{load}}{k_n(V_{OH}-V_{T,n})}\ln\left(\frac{2(V_{OH}-V_{T,n})-V_{50\%}}{V_{50\%}}\right.$$

$$\tau_{PHL} = \frac{C_{load}}{k_n(V_{OH}-V_{T,n})}\left[\frac{2V_{T,n}}{V_{OH}-V_{T,n}} + \ln\left(\frac{4(V_{OH}-V_{T,n})}{V_{OH}+V_{OL}} - 1\right)\right.$$

$$\tau_{PHL} = \frac{C_{load}}{k_n(V_{DD}-V_{T,n})}\left[\frac{2V_{T,n}}{V_{DD}-V_{T,n}} + \ln\left(\frac{4(V_{DD}-V_{T,n})}{V_{DD}} - 1\right)\right]$$

Example 6.1

Consider the CMOS inverter circuit shown in Fig. 6.2, with $V_{DD} = 3.3$ V. The *I–V* characteristics of the nMOS transistor are specified as follows: when $V_{GS} = 3.3$ V, the drain current reaches its saturation level $I_{sat} = 2$ mA for $V_{DS} \geq 2.5$ V. Assume that the input signal applied to the gate is a step pulse that switches instantaneously from 0 V to 3.3 V. Using the data above, calculate the delay time necessary for the output to fall from its initial value of 3.3 V to 1.65 V, assuming an output load capacitance of 300 fF.

For the solution, consider the simplified pull-down circuit shown in Fig. 6.5. We will assume that the nMOS transistor operates in saturation from $t = 0$ to $t = t_1' = t_{sat}$, and that it will operate in the linear region from $t = t_1' = t_{sat}$ to $t = t_2 = t_{delay}$. We can also deduce from the *I–V* characteristics that $V_{T,n} = 0.8$ V, since the nMOS transistor enters saturation when $V_{DS} \geq V_{GS} - V_{T,n}$. The voltage V_{GS} is equal to 3.3 V for $t \geq 0$.

The current equation for the saturation region can be written as

$$C\frac{dV_{out}}{dt} = -I_D = -I_{sat} = -\frac{1}{2}k_n(V_{OH} - V_{T,n})^2$$

We can calculate the amount of time in which the nMOS transistor operates in saturation (t_{sat}), by integrating this equation.

$$\int_{t=0}^{t=t_{sat}} dt = -\int_{V_{out}=3.3}^{V_{out}=2.5} \frac{C}{I_D}\, dV_{out}$$

$$t_{sat} = \frac{V_{T,n}C}{I_{sat}} = \frac{0.8\,\text{V} \cdot 300\,\text{fF}}{2\,\text{mA}} = 120\,[\text{ps}]$$

The transconductance k_n of the nMOS transistor can be found as follows:

$$k_n = \frac{2I_{sat}}{(V_{OH} - V_{T,n})^2} = \frac{2 \times 2 \times 10^{-3}}{(3.3 - 0.8)^2} = 0.640 \times 10^{-3}\ [\text{A/V}^2]$$

Now, the current equation for the linear operating region is

$$C\frac{dV_{out}}{dt} = -I_D = -\frac{1}{2}k_n\left[2(V_{OH} - V_{T,n})V_{out} - V_{out}^2\right]$$

Integrating this differential equation between the two voltage boundary conditions yields the time in which the nMOS transistor operates in the linear region during this transition.

$$\int_{t=t_{sat}}^{t=t_{delay}} dt = -2C \int_{V_{out}=2.5}^{V_{out}=1.65} \frac{dV_{out}}{k_n\left[2(V_{OH} - V_{T,n})V_{out} - V_{out}^2\right]}$$

$$t_{delay} - t_{sat} = -\frac{C}{k_n}\frac{1}{(V_{OH} - V_{T,n})} \ln\left(\frac{V_{out}}{2(V_{OH} - V_{T,n}) - V_{out}}\right)\Bigg|_{V_{out}=2.5}^{V_{out}=1.65}$$

$$= \frac{C}{k_n}\frac{1}{(V_{OH} - V_{T,n})} \times \left[\ln\left(\frac{2(V_{OH} - V_{T,n}) - V_{1.65}}{V_{1.65}}\right) - \ln\left(\frac{2(V_{OH} - V_{T,n}) - V_{2.5}}{V_{2.5}}\right)\right]$$

$$= \frac{0.3 \times 10^{-12}}{0.640 \times 10^{-3}}\frac{1}{(3.3 - 0.8)}\left[\ln\left(\frac{5 - 1.65}{1.65}\right) - \ln\left(\frac{5 - 2.5}{2.5}\right)\right]$$

$$= 133\ [\text{ps}]$$

Thus, the total delay time is found to be

$$t_{delay} = 120 + 133 = 253\ [\text{ps}]$$

Note that t_{delay} corresponds to the propagation delay time τ_{PHL} for falling output.

Example 6.2

For the CMOS inverter shown in Fig. 6.2 with a power supply voltage of $V_{DD} = 5$ V, determine the fall time τ_{fall}, which is defined as the time elapsed between the time point at which $V_{out} = V_{90\%} = 4.5$ V and the time point at which $V_{out} = V_{10\%} = 0.5$ V. Use both the average-current method and the differential equation method for calculating τ_{fall}. The output load capacitance is 1 pF. The nMOS transistor parameters are given as

$$\mu_n C_{ox} = 20\,\mu\text{A/V}^2$$
$$(W/L)_n = 10$$
$$V_{T,n} = 1.0\text{ V}$$

Using a simple expression similar to (6.10), we can determine the average capacitor current during the charge-down event described earlier.

$$I_{avg} = \frac{1}{2}[I(V_{in} = 5\text{ V}, V_{out} = 4.5\text{ V}) + I(V_{in} = 5\text{ V}, V_{out} = 0.5\text{ V})]$$
$$= \frac{1}{2}\left[\frac{1}{2}k_n(V_{in} - V_{T,n})^2 + \frac{1}{2}k_n\left(2(V_{in} - V_{T,n})V_{out} - V_{out}^2\right)\right]$$
$$= \frac{1}{2}\cdot\frac{1}{2}\cdot 20\times 10^{-6}\cdot 10[(5-1)^2 + (2(5-1)0.5 - 0.5^2)] = 0.9875\text{ [mA]}$$

The fall time is then found as

$$\tau_{fall} = \frac{C\cdot\Delta V}{I_{avg}} = \frac{1\times 10^{-12}(4.5-0.5)}{0.9875\times 10^{-3}} = 4.05\times 10^{-9}\text{ [s]} = 4.05\text{ [ns]}$$

Now, we will recalculate the fall time using the differential equation approach. The nMOS transistor operates in the saturation region for $4.0\text{ V} \le V_{out} \le 4.5$ V. Writing the current equation for the saturation region, we obtain

$$C\frac{dV_{out}}{dt} = -\frac{1}{2}k_n(V_{in} - V_{T,n})^2, \qquad \text{where} \quad k_n = \mu_n C_{ox}\left(\frac{W}{L}\right)_n$$

$$\frac{dV_{out}}{dt} = \frac{-20\times 10^{-6}\cdot 10\cdot(5-1)^2}{2\cdot 1\times 10^{-12}} = -1.6\times 10^9\text{ [V/s]}$$

Integrating this equation, we obtain the delay component during which the nMOS transistor operates in the linear region.

$$\int_{t=t_{sat}}^{t=t_{delay}} dt = -2C\int_{V_{out}=4}^{V_{out}=0.5}\frac{dV_{out}}{k_n\left[2(V_{in} - V_{T,n})V_{out} - V_{out}^2\right]}$$

$$\tau_{fall} - t_{sat} = \frac{C}{k_n}\frac{1}{(V_{in} - V_{T,n})}$$
$$\times\left[\ln\left(\frac{2(V_{in} - V_{T,n}) - V_{0.5}}{V_{0.5}}\right) - \ln\left(\frac{2(V_{in} - V_{T,n}) - V_{4.0}}{V_{4.0}}\right)\right]$$

$$= \frac{1\times 10^{-12}}{20\times 10^{-6}\cdot 10\cdot 4}\left[\ln\left(\frac{8-0.5}{0.5}\right) - \ln\left(\frac{8-4}{4}\right)\right]$$

$$= 3.385\times 10^{-9}\text{ [s]} = 3.385\text{ [ns]}$$

Thus, the fall time of the CMOS inverter is found as follows:

$$\tau_{fall} = 3.6975\text{ [ns]}$$

Integrating this simple expression yields the time during which the nMOS transistor operates in saturation.

$$\int_{t=0}^{t=t_{sat}} dt = -\frac{1}{1.6\times 10^9}\int_{V_{out}=4.5}^{V_{out}=4} dV_{out}$$

$$t_{sat} = \frac{0.5}{1.6\times 10^9} = 0.3125\times 10^{-9}\text{ [s]} = 0.3125\text{ [ns]}$$

The nMOS transistor operates in the linear region for $0.5\text{ V} \le V_{out} \le 4.0$ V. The current equation for this operating region is written as follows:

$$C\frac{dV_{out}}{dt} = -\frac{1}{2}k_n\left[2(V_{in} - V_{T,n})V_{out} - V_{out}^2\right]$$

Calculation of delay time (2)

In a CMOS inverter, the charge - up event of the output load capacitanc e for falling input tran sition is completely analogous to the charge - down event for rising input

$$\tau_{PLH} = \frac{C_{load}}{k_p(V_{OH} - V_{OL} - |V_{T,p}|)} \times \left[\frac{2|V_{T,P}|}{V_{OH} - V_{OL} - |V_{T,p}|} + \ln\left(\frac{2(V_{OH} - V_{OL} - |V_{T,p}|)}{V_{OH} - V_{50\%}} \right) - 1 \right]$$

$$\tau_{PLH} = \frac{C_{load}}{k_p(V_{DD} - |V_{T,p}|)} \times \left[\frac{|V_{T,P}|}{V_{DD} - |V_{T,p}|} + \ln\left(\frac{4(V_{DD} - |V_{T,p}|)}{V_{DD}} \right) - 1 \right]$$

the sufficient conditions for balanced propagatio n delays, i.e. for $\tau_{PHL} = \tau_{PLH}$

$\Rightarrow V_{T,n} = |V_{T,p}|$ *and* $k_n = k_p \left(or\ W_p/W_n = \mu_n/\mu_p \right)$

When the input volt age switches from high to low, nMOS off, pMOS on charging load $C_{load} \frac{dV_{out}}{dt} = i_{D,load}(V_{out})$, note the pMOS initially in saturation , and enter linear when the output vol tage is rises above $\left(V_{DD} + V_{T,load} \right)$

$$i_{D,load} = \frac{k_{n,load}}{2}(|V_{T,load}|) \quad \text{for } V_{out} \le V_{DD} - |V_{T,load}|$$

$$i_{D,load} = \frac{k_{n,load}}{2}\left[2|V_{T,load}|(V_{DD} - V_{out}) - (V_{DD} - V_{out})^2 \right] \text{for } V_{out} > V_{DD} - |V_{T,load}|$$

$$\tau_{PLH} = C_{load}\left[\int_{V_{out}=V_{OL}}^{V_{out}=V_{DD}-|V_{T,load}|} \left(\frac{dV_{out}}{i_{D,load}(sat)} \right) + \int_{V_{out}=V_{DD}-|V_{T,load}|}^{V_{out}=V_{50\%}} \left(\frac{dV_{out}}{i_{D,load}(linear)} \right) \right]$$

$$\tau_{PLH} = \frac{C_{load}}{k_{n,load}|V_{T,load}|}\left[\frac{2(V_{DD} - |VT,load| - V_{OL})}{|V_{T,load}|} + \ln\left(\frac{2|V_{T,load}| - (V_{DD} - V_{50\%})}{V_{DD} - V_{50\%}} \right) \right]$$

$$\tau_{PHL}(actual) = \sqrt{\tau_{PHL}{}^2(stepinput) + \left(\frac{\tau_r}{2} \right)^2}$$

$$\tau_{PLH}(actual) = \sqrt{\tau_{PLH}{}^2(stepinput) + \left(\frac{\tau_f}{2} \right)^2}$$

Calculation of delay time (3)

- Considering the input voltage waveform is not an ideal (step) pulse waveform, but has finite rise and fall times
  - Using an empirical expression as 6.29, 6.30
- The former expression based on the gradual channel approximation
  - Can still be used for sub-micron MOS transistors with proper parameter adjustments
  - Yet , the current driving capability of sub-micron transistors is significantly reduced as a result of channel velocity saturation
    - (W/L)-ratio
    - In deep-sub-micron nMOS ⇨saturation current no longer$\propto(V_{GS}-V_T)^2$
      - $I_{sat}=\kappa W_n(V_{GS}-V_T)$
      - $\tau_{PHL}\approx(C_{load}V_{50\%})/I_{sat}=[C_{load}(V_{DD}/2)]/\ \kappa W_n(V_{GS}-V_T)$
      - The propagation delay has only a weak dependence od the power supply
  - Better estimate can be obtain by using an accurate short-channel MOSFET model

Inverter design with delay constrains

- The load capacitance C_{load} consist of
  - Intrinsic components ⇨ parasitic drain capacitances which depend on transistor dimensions
  - Extrinsic component ⇨ interconnect/wiring capacitance and fan-out capacitance
- If C_{load} mainly consists of extrinsic components, and if this overall load capacitance can be estimated accurately and independently of the transistor dimensions
  - Given a required (target) delay value of τ^*_{PHL}
  - The (W/L)-ratio can be found as

$$\left(\frac{W_n}{L_n}\right) = \frac{C_{load}}{\tau_{PHL}^{*}\,\mu_n C_{ox}(V_{DD}-V_{T,n})}\left[\frac{2V_{T,n}}{V_{DD}-V_{T,n}} + \ln\left(\frac{4(V_{DD}-V_{T,n})}{V_{DD}} - 1\right)\right]$$

$$\left(\frac{W_p}{L_p}\right) = \frac{C_{load}}{\tau_{PLH}^{*}\,\mu_p C_{ox}(V_{DD}-V_{T,p})}\left[\frac{2\,|V_{T,p}|}{V_{DD}-|V_{T,p}|} + \ln\left(\frac{4(V_{DD}-|V_{T,p}|)}{V_{DD}} - 1\right)\right]$$

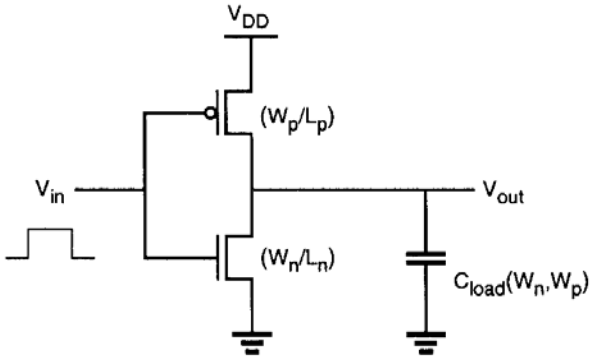


Figure 6.7 General circuit structure considered in the inverter design problem.

Example 6.3

A company has access to a CMOS fabrication process with the device parameters listed below.

$\mu_n C_{ox} = 120\,\mu\text{A/V}^2$
$\mu_p C_{ox} = 60\,\mu\text{A/V}^2$
$L = 0.6\,\mu\text{m}$ for both nMOS and pMOS devices
$V_{T0,n} = 0.8$ V
$V_{T0,p} = -1.0$ V
$W_{min} = 1.2\,\mu\text{m}$

Design a CMOS inverter by determining the channel widths W_n and W_p of the nMOS and pMOS transistors, to meet the following performance specifications.

- $V_{th} = 1.5$ V for $V_{DD} = 3$ V,
- Propagation delay times $\tau^*_{PHL} \leq 0.2$ ns and $\tau^*_{PLH} \leq 0.15$ ns,
- A falling delay of 0.35 ns for an output transition from 2 V to 0.5 V, assuming a combined output load capacitance of 300 fF and ideal step input.

We start our design by satisfying the time delay constraints. First, the *minimum* (W/L) ratios of the nMOS and pMOS transistors which are dictated by the propagation delay constraints can be found using (6.33) and (6.34), as follows.

$$\left(\frac{W_n}{L_n}\right) = \frac{C_{load}}{\tau^*_{PHL}\mu_n C_{ox}(V_{DD} - V_{T,n})}\left[\frac{2V_{T,n}}{V_{DD} - V_{T,n}} + \ln\left(\frac{4(V_{DD} - V_{T,n})}{V_{DD}} - 1\right)\right]$$

$$= \frac{300 \times 10^{-15}}{0.2 \times 10^{-9} \cdot 120 \times 10^{-6} \cdot (3 - 0.8)}\left[\frac{2 \cdot 0.8}{3 - 0.8} + \ln\left(\frac{4(3 - 0.8)}{3} - 1\right)\right]$$

$$= 7.9$$

$$\left(\frac{W_p}{L_p}\right) = \frac{C_{load}}{\tau^*_{PLH}\mu_p C_{ox}(V_{DD} - |V_{T,p}|)}$$

$$\times\left[\frac{2|V_{T,p}|}{V_{DD} - |V_{T,p}|} + \ln\left(\frac{4(V_{DD} - |V_{T,p}|)}{V_{DD}} - 1\right)\right]$$

$$= \frac{300 \times 10^{-15}}{0.15 \times 10^{-9} \cdot 60 \times 10^{-6} \cdot (3 - 1)}\left[\frac{2 \cdot 1}{3 - 1} + \ln\left(\frac{4(3 - 1)}{3} - 1\right)\right]$$

$$= 25.2$$

During the falling output transition (from 2 V to 0.5 V), the nMOS transistor of the CMOS inverter will operate entirely in the linear region. The current equation of the nMOS transistor in this region is

$$C_{load}\frac{dV_{out}}{dt} = -\frac{1}{2}\mu_n C_{ox}\frac{W_n}{L_n}\left[2(V_{OH} - V_{T0,n})V_{out} - V_{out}^2\right]$$

By integrating this expression, we obtain the following relationship.

$$t_{delay} = 0.35 \times 10^{-9} = -2C_{load}\int_{V_{out}=2}^{V_{out}=0.5}\frac{dV_{out}}{\mu_n C_{ox}\frac{W_n}{L_n}\left[2(V_{OH} - V_{T0,n})V_{out} - V_{out}^2\right]}$$

$$t_{delay} = \frac{-C_{load}}{\mu_n C_{ox}\frac{W_n}{L_n}}\frac{1}{(V_{OH} - V_{T0,n})}\ln\left(\frac{V_{out}}{2(V_{OH} - V_{T0,n}) - V_{out}}\right)\Bigg|_2^{0.5}$$

$$t_{delay} = \frac{-C_{load}}{\mu_n C_{ox}\left(\frac{W_n}{L_n}\right)}\frac{1}{(3 - 0.8)}\left[\ln\left(\frac{0.5}{2(3 - 0.8) - 0.5}\right) - \ln\left(\frac{2}{2(3 - 0.8) - 2}\right)\right]$$

$$0.35 \times 10^{-9} = \frac{-300 \times 10^{-15}}{120 \times 10^{-6}\left(\frac{W_n}{L_n}\right)2.2}[-2.054 + 0.182]$$

Now we solve this equation for the nMOS transistor (W/L) ratio:

$$\left(\frac{W_n}{L_n}\right) = 6.1$$

Notice that this ratio is *smaller* than the (W/L)-ratio found from the propagation delay constraint. Thus, we take the larger ratio which will satisfy both timing constraints, and determine the size of the nMOS transistor as $W_n = 4.7\,\mu$m, for the given $L_n = 0.6\,\mu$m. Next, the logic threshold constraint of $V_{th} = 1.5$ V will help determine the pMOS transistor dimensions. Using (5.87) for the logic threshold voltage of the CMOS inverter,

$$V_{th} = \frac{V_{T0,n} + \sqrt{\frac{1}{k_R}}(V_{DD} + V_{T0,p})}{1 + \sqrt{\frac{1}{k_R}}} = 1.5$$

we find that the ratio k_R which satisfies this design constraint is equal to 0.51. This value can now be used to calculate the (W/L)-ratio of the pMOS transistor, as follows.

$$k_R = \frac{\mu_n C_{ox}\left(\frac{W_n}{L_n}\right)}{\mu_p C_{ox}\left(\frac{W_p}{L_p}\right)} = \frac{120 \times 10^{-6}\,(7.9)}{60 \times 10^{-6}\left(\frac{W_p}{L_p}\right)} = 0.51$$

$$\left(\frac{W_p}{L_p}\right) = 31$$

Note that this ratio is *larger* than the one found from the propagation delay constraint earlier. Since the larger ratio will satisfy both the timing constraint and the V_{th}-constraint, we determine the pMOS transistor size as $W_p = 18.6\,\mu$m, for the given $L_p = 0.6\,\mu$m.

Inverter design with delay constrains

If Cload have to take into accunt that the intrinsic component

$$C_{load} = C_{gd,n}(W_n) + C_{gd,p}(W_p) + C_{db,n}(W_n) + C_{db,p}(W_p) + C_{\text{int}} + C_g = f(W_n, W_p)$$

the fan - out capacotance C_g is also a function of the device dimensions in the next - stage gates

Condering the simplified layout in Fig. 6.8 $\Rightarrow$

The relatively small gate - to - drain capacitances $C_{gd,n}$ and $C_{gd,p}$ will be neglected in the analysis

The drain parasitic capacitance are :

$$C_{db,n} = W_n D_{drain} C_{j0,n} K_{eq,n} + 2(W_n + D_{drain}) C_{jsw,n} K_{eq,n}$$

$$C_{db,p} = W_p D_{drain} C_{j0,p} K_{eq,p} + 2(W_p + D_{drain}) C_{jsw,p} K_{eq,p}$$

$$C_{load} = (W_n C_{j0,n} K_{eq,n} + W_p C_{j0,p} K_{eq,p}) D_{drain} + 2(W_n + D_{drain}) C_{jsw,n} K_{eq,n} + 2(W_p + D_{drain}) C_{jsw,p} K_{eq,p} + C_{\text{int}} + C_g$$

The total capacitive load can be expressed as :

$$C_{load} = \alpha_0 + \alpha_n W_n + \alpha_p W_p$$

where $\alpha_0 = 2D_{drain}(C_{jsw,n} K_{eq,n} + C_{jsw,p} K_{eq,p}) + C_{\text{int}} + C_g$

$$\alpha_n = K_{eq,n}(C_{j0,n} D_{drain} + 2C_{jsw,n})$$

$$\alpha_p = K_{eq,p}(C_{j0,p} D_{drain} + 2C_{jsw,p})$$

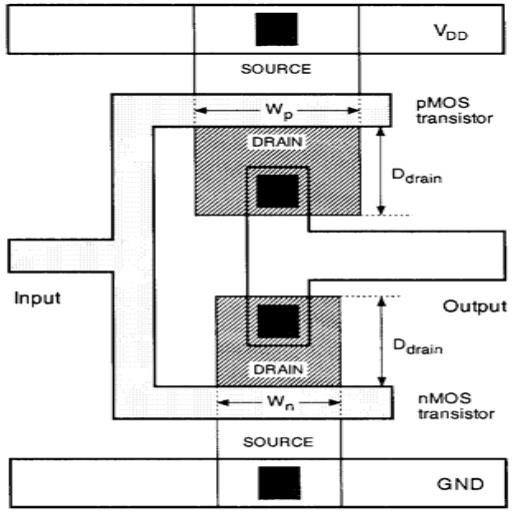


Figure 6.8 Simplified CMOS inverter mask layout used for delay analysis.

Inverter design with delay constrains

The propagation delay :

$$\tau_{PHL} = \left(\frac{\alpha_0 + \alpha_n W_n + \alpha_p W_p}{W_n}\right) \times \left(\frac{L_n}{\mu_n C_{ox}(V_{DD} - V_{T,n})}\right) \times \left[\frac{2V_{T,n}}{V_{DD} + V_{T,n}} + \ln\left(\frac{4(V_{DD} - V_{T,n})}{V_{DD}} - 1\right)\right]$$

$$\tau_{PLH} = \left(\frac{\alpha_0 + \alpha_n W_n + \alpha_p W_p}{W_n}\right) \times \left(\frac{L_p}{\mu_p C_{ox}(V_{DD} - |V_{T,p}|)}\right) \times \left[\frac{2|V_{T,p}|}{V_{DD} - |V_{T,p}|} + \ln\left(\frac{4(V_{DD} - |V_{T,p}|)}{V_{DD}} - 1\right)\right]$$

Note that the channel lengths L_n and L_p are usually fixed and equal to each other

the transistor *aspect ratio* be defined as : $R(\text{aspect ratio}) \equiv \frac{W_p}{W_n}$

$$\tau_{PHL} = \Gamma_n\left(\frac{\alpha_0 + (\alpha_n + R\alpha_p)W_n}{W_n}\right), \quad \tau_{PLH} = \Gamma_p\left(\frac{\alpha_0 + (\frac{\alpha_n}{R} + \alpha_p)W_p}{W_p}\right)$$

where $\Gamma_n = \left(\frac{L_n}{\mu_n C_{ox}(V_{DD} - V_{T,n})}\right) \times \left[\frac{2V_{T,n}}{V_{DD} - V_{T,n}} + \ln\left(\frac{4(V_{DD} - V_{T,n})}{V_{DD}} - 1\right)\right]$

$$\Gamma_p = \left(\frac{L_p}{\mu_p C_{ox}(V_{DD} - |V_{T,p}|)}\right) \times \left[\frac{2|V_{T,P}|}{V_{DD} - |V_{T,p}|} + \ln\left(\frac{4(V_{DD} - |V_{T,P}|)}{V_{DD}} - 1\right)\right]$$

Given target delay value τ^*_{PHL} and τ^*_{PLH} the minimum channel widths of the nMOS transistor and the pMOS transistor which satisfy these delay constraints can be calculated from (6.46a) and (6.46b), by solving for W_n and W_p, respectively.

Increasing Wn and Wp to reduce the propagation delay times will have a dimishing influence upon delay beyond certain values,

$$\tau^{\text{limit}}_{\text{PHL}} = \Gamma_n(\alpha_n + R\alpha_p), \quad \tau^{\text{limit}}_{\text{PLH}} = \Gamma_p\left(\frac{\alpha_n}{R} + \alpha_p\right)$$

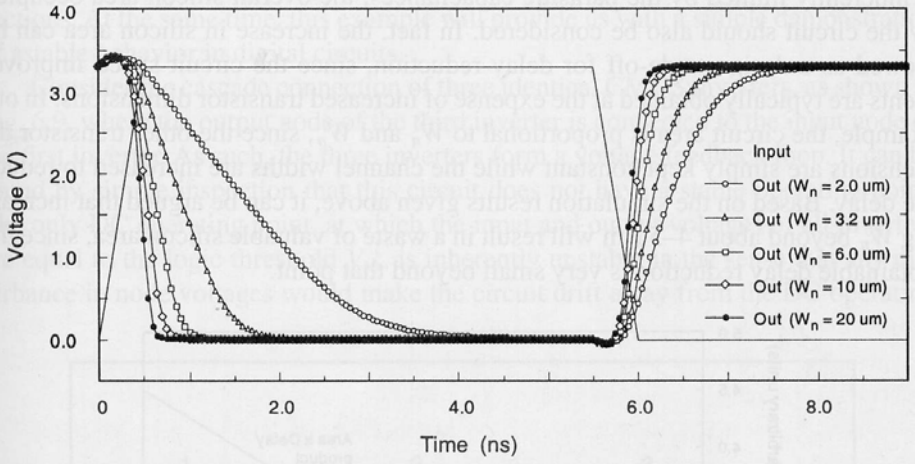
The propagation delay cannot be reduced beyond these limit values, which are dictated by technology - related parameters .

The propagation delay time is independent of the extrinsic capacitance component, C_{int} and C_g

Example 6.4

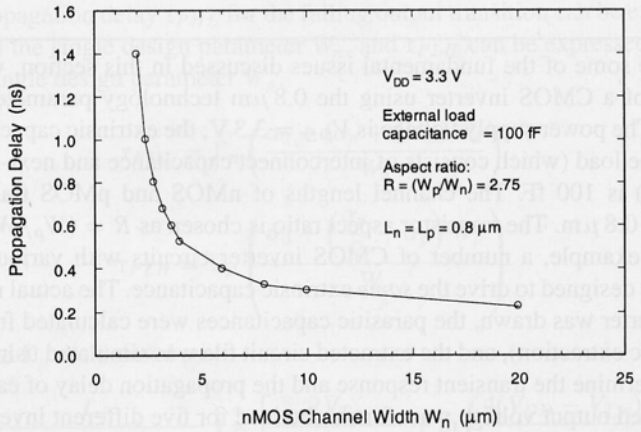
To illustrate some of the fundamental issues discussed in this section, we consider the design of a CMOS inverter using the 0.8 μm technology parameters given in Chapter 4. The power supply voltage is $V_{DD} = 3.3$ V; the extrinsic capacitance component of the load (which consists of interconnect capacitance and next-stage fan-in capacitance) is 100 fF. The channel lengths of nMOS and pMOS transistors are $L_n = L_p = 0.8\,\mu$m. The transistor aspect ratio is chosen as $R = (W_p/W_n) = 2.75$.

In this example, a number of CMOS inverter circuits with various transistor widths were designed to drive the *same* extrinsic capacitance. The actual mask layout of each inverter was drawn, the parasitic capacitances were calculated from the layout (parasitic extraction), and the extracted circuit file was simulated using SPICE in order to determine the transient response and the propagation delay of each inverter. The simulated output voltage waveforms obtained for five different inverter designs are shown below.



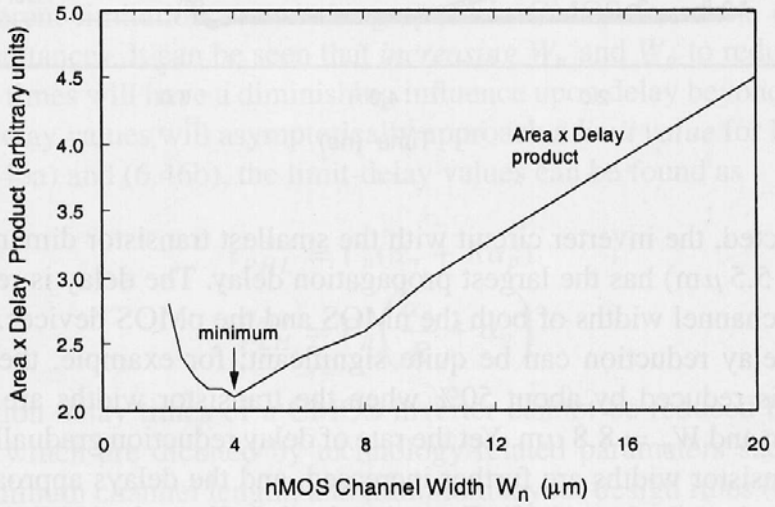
As expected, the inverter circuit with the smallest transistor dimensions ($W_n = 2\,\mu$m, $W_p = 5.5\,\mu$m) has the largest propagation delay. The delay is reduced by increasing the channel widths of both the nMOS and the pMOS devices. Initially, the amount of delay reduction can be quite significant; for example, the propagation delay τ_{PHL} is reduced by about 50% when the transistor widths are increased to $W_n = 3.2\,\mu$m and $W_p = 8.8\,\mu$m. Yet the rate of delay reduction gradually diminishes when the transistor widths are further increased, and the delays approach limit values which are described by (6.48). For example, the effect of a 100% width increase from $W_n = 10\,\mu$m to $W_n = 20\,\mu$m is almost negligible, due to the increased drain parasitic capacitances of both transistors, as explained in the previous discussion.

In the following figure, the falling-output propagation delay τ_{PHL} (obtained from SPICE simulation) is plotted as a function of the nMOS channel width. The delay asymptotically approaches a limit value of about 0.2 ns, which is mainly determined by technology-specific parameters, independent of the extrinsic capacitance component.



Example 6.4

In addition to the fact that the influence of device sizing upon propagation delay is inherently limited by the parasitic capacitances, the overall silicon area occupied by the circuit should also be considered. In fact, the increase in silicon area can be viewed as a design trade-off for delay reduction, since the circuit speed improvements are typically obtained at the expense of increased transistor dimensions. In our example, the circuit area is proportional to W_n and W_p, since the other transistor dimensions are simply kept constant while the channel widths are increased to reduce the delay. Based on the simulation results given above, it can be argued that increasing W_n beyond about 4–5 μm will result in a waste of valuable silicon area, since the obtainable delay reduction is very small beyond that point.



A practical measure used for quantifying *design quality* is the (Area × Delay) product, which takes into account the silicon-area cost of transistor sizing for delay reduction. While the propagation delay asymptotically approaches a limit value for increasing channel widths, the (Area × Delay) product exhibits a clear minimum around $W_n = 4\,\mu$m, indicating the optimum choice both in terms of speed and overall silicon area.

CMOS ring oscillator circuit

- This circuit does not have a stable operating point
- The only DC operating point:
  - the input and output voltages of all inverters are equal to the logic threshold V_{th} (unstable)
- A closed-loop cascade connection of any odd number of inverter will display astable behavior
  - will oscillate once any of the inverter input or output voltages deviate from the unstable operating point, V_{th}
  - V_1, $V_{OL} \rightarrow V_{OH}$ ⇨ trigger V_2 to fall, $V_{OH} \rightarrow V_{OL}$, difference between the $V_{50\%}$-crossing times of V_1 and V_2, τ_{PHL2} ⇨ trigger V_3 to rise, $V_{OL} \rightarrow V_{OH}$, difference between the $V_{50\%}$-crossing times of V_2 and V_3, τ_{PHL3}.....
  - $T = \tau_{PHL1} + \tau_{PHL1} + \tau_{PHL2} + \tau_{PHL2} + \tau_{PHL3} + \tau_{PHL3} = 6\tau_P$
  - $f = 1/T = 1/(2n\tau_P)$, $\tau_P = 1/2nf$

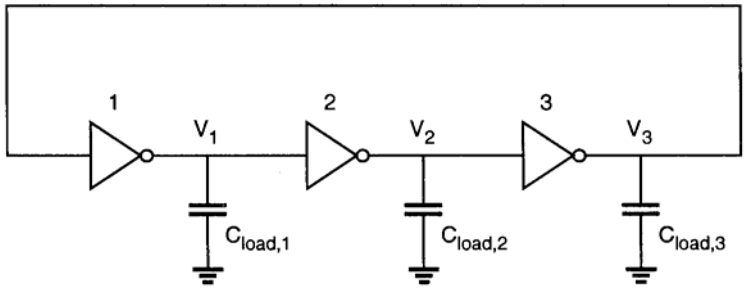


Figure 6.9 Three-stage ring oscillator circuit consisting of identical inverters.

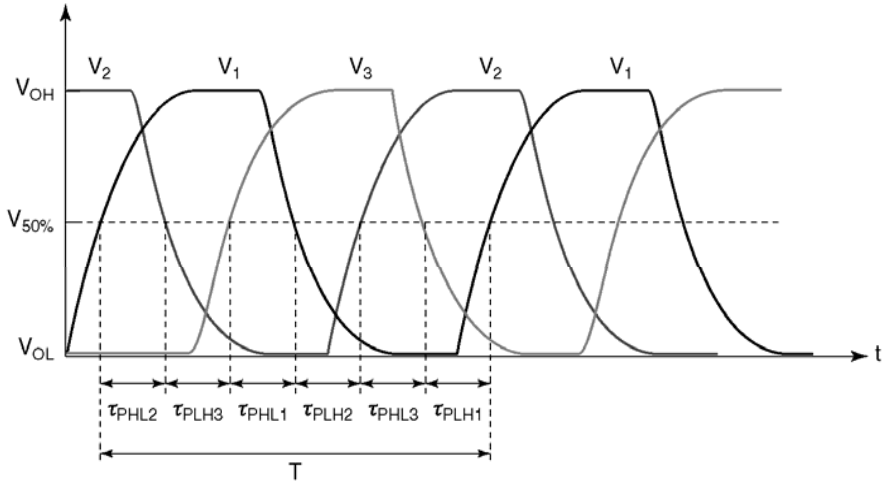


Figure 6.10 Typical voltage waveforms of the three inverters shown in Fig. 6.9.

Estimation of interconnect parasitics

- The load
  - Classical approach ⇨capacitive and lumped
    - Internal parasitic capacitance of the transistor
    - Interconnect (line) capacitances
    - Input capacitances of the fan-out gates
- Now, the interconnect line itself
  - Three dimensional structure in metal and/or polysilicon
    - Non-negligible resistance
    - The (length/width) ratio of the wire ⇨ distributed ⇨ making the interconnect a true transmission line
    - An interconnect is rarely isolated from other influence

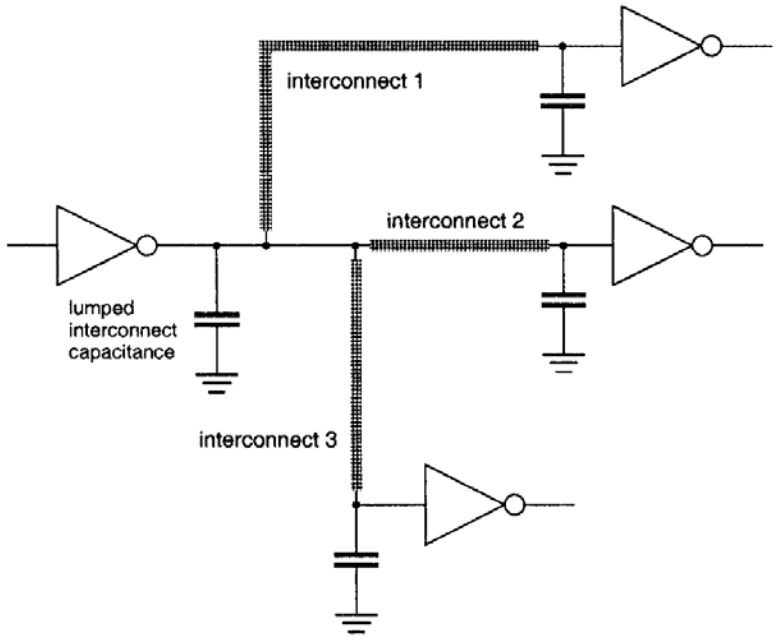


Figure 6.11 An inverter driving three other inverters over interconnection lines.

Estimation of interconnect parasitics

- If the time of flight across the interconnection line is much shorter than the signal rise/fall times
  - The wire can be modeled as a capacitive load, or as a lumped or distributed RC network
- If the interconnection lines are sufficient long and the rise times of the signal comparable to …
  - The inductance becomes important
  - Modeled as transmission lines
  - $$\tau_{rise}(\tau_{fall}) < 2.5\times\left(\frac{l}{v}\right) \Rightarrow \{\text{transmission} - \text{line modeling}\}$$

    $$2.5\times\left(\frac{l}{v}\right) < \tau_{rise}(\tau_{fall}) < 5\times\left(\frac{l}{v}\right) \Rightarrow \begin{Bmatrix}\text{either transmission - line}\\ \text{or lumped modeling}\end{Bmatrix}$$

    $$\tau_{rise}(\tau_{fall}) > 5\times\left(\frac{l}{v}\right) \Rightarrow \{\text{lumped modeling}\}$$

    Here, l is the interconnect line length, and v is the propagation speed
  - The longest wire on a VLSI chip (2cm)⇨fright time≅133ps, shorter than rise/fall time⇨ capacitive or RC model
  - 10 cm multi-chip module ⇨1ns, the same order as rise/fall time ⇨considering RLCG

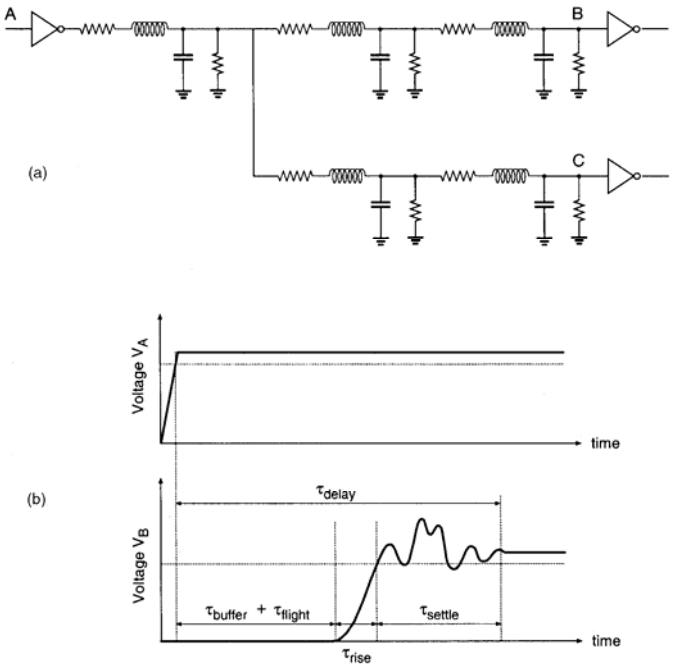


Figure 6.12 (a) An RLCG interconnection tree. (b) Typical signal waveforms at the nodes A and B, showing the signal delay and the various delay components.

The transmission line effect

- IN CMOS VLSI chips
  - Not serious concern
  - The gate delay due to capacitive load component dominated the line delay
- The sub-micron design rules
  - The intrinsic gate delay tend to decrease significantly
  - The overall chip size and the worse-case line length on a chip tend to increase
    - Mainly due to increasing chip complexity
    - The widths of metal lines shrink while thickness increase
      - The transmission line effects and signal coupling between neighboring lines become even more pronounced
- To optimize a system for speed, chip designer must have reliable and efficient means for
  - Estimating the interconnect parasitics in a large chip
  - Simulating the transient effect

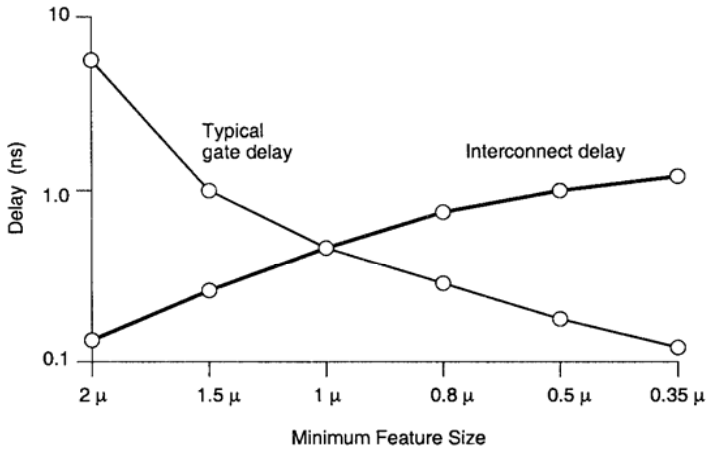


Figure 6.13 Interconnect delay dominates gate delay in sub-micron CMOS technologies.

Interconnection delay

- The hierarchical structure of most VLSI design
  - Chip
  - Modules
    - Inter-module connection ⇨ longer
  - Logic gates, transistors
    - Intra-module connection ⇨shorter

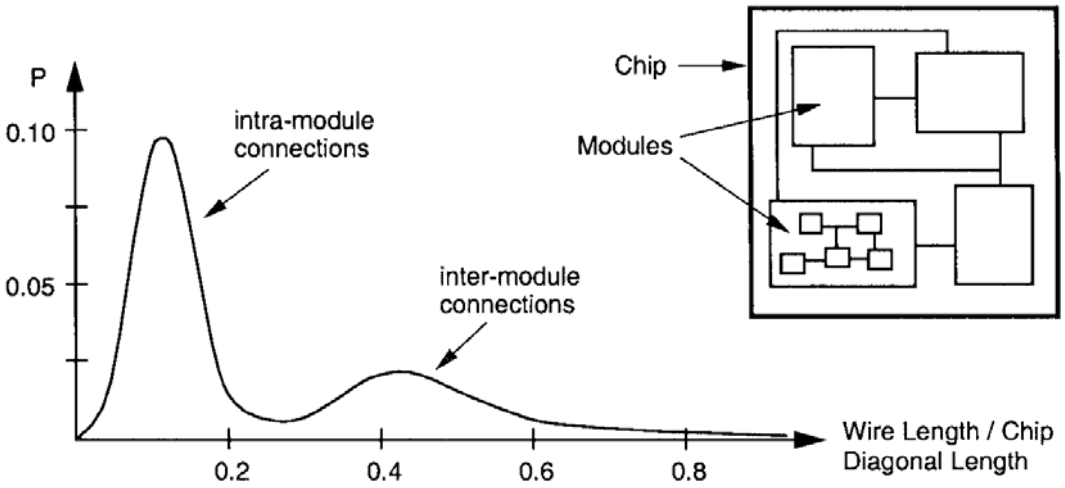


Figure 6.14 Statistical distribution of interconnection length on a typical chip.

Interconnect capacitance estimation

- A complicated task
- Fringing-field factor FF=C_{total}

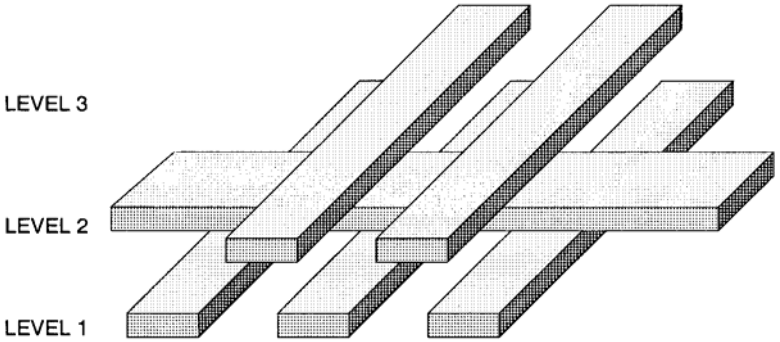


Figure 6.15 An example of six interconnect lines running on three different levels.

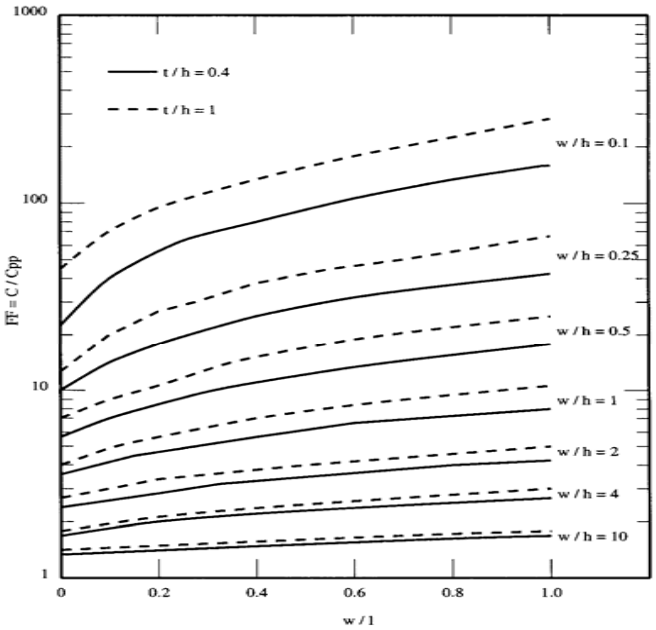


Figure 6.18 Variation of the fringing-field factor with the interconnect geometry.

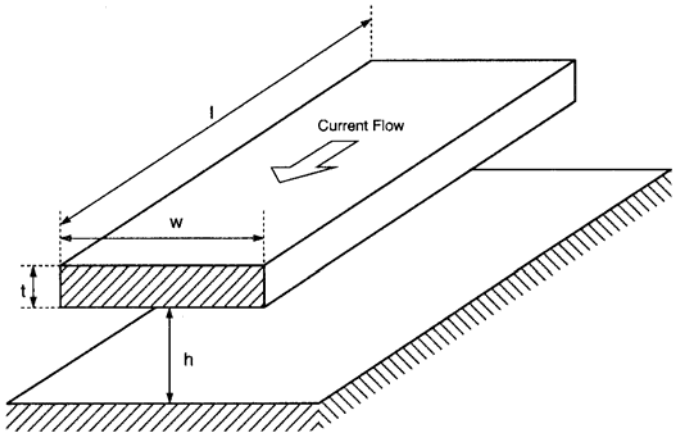


Figure 6.16 Interconnect segment running parallel to the surface, which is used for parasitic resistance and capacitance estimations.

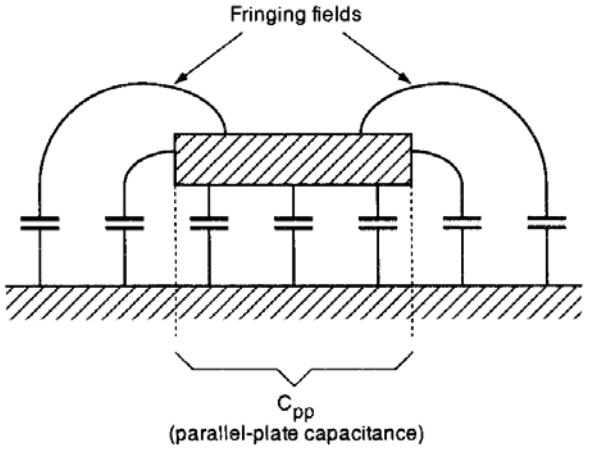


Figure 6.17 Influence of fringing electric fields upon the parasitic wire capacitance.

Estimation of interconnection capacitance

- The formulas provide accurate approximation of the parasitic capacitance values to within 10% error, even for very small values of (w/h) and (t/h)
  - The linear dash-dotted line ⇨parallel-plate cap.
  - W/T decreases ⇨cap. Decreases
    - Level off at approximately 1pF/cm, when the wire width is approximately equal to insulator thickness

$$C = \varepsilon \left[\frac{\left(w - \frac{t}{2} \right)}{h} + \frac{2\pi}{\ln\left(1 + \frac{2h}{t} + \sqrt{\frac{2h}{t}\left(\frac{2h}{t} + 2 \right)} \right)} \right] \quad for \ \mathrm{w} \geq \frac{\mathrm{t}}{2}$$

$$C = \varepsilon \left[\frac{w}{h} + \frac{\pi\left(1 - 0.0543 \cdot \frac{t}{2h} \right)}{\ln\left(1 + \frac{2h}{t} + \sqrt{\frac{2h}{t}\left(\frac{2h}{t} + 2 \right)} \right)} + 1.47 \right] \quad for \ \mathrm{w} < \frac{\mathrm{t}}{2}$$

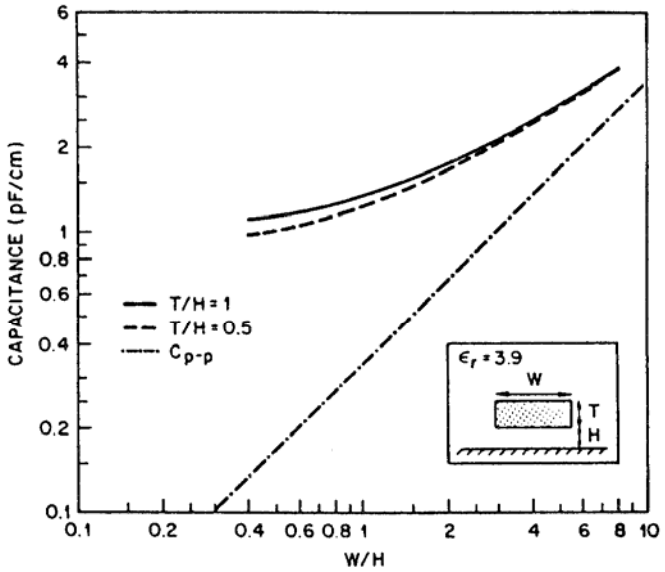


Figure 6.19 Capacitance of a single interconnect, as a function of (*w/h*) and (*t/h*).

Capacitance coupling

- Considering the interconnection line is not completely isolated from the surrounding structures, but is coupled with other lines running in parallel
  - The total parasitic capacitance increased by
    - Fringing-field effects
    - Capacitive coupling between the lines
      - When the thickness of the wire is comparable to its width ⇨ coupling capacitance↑
      - Signal crosstalk
        - » Transitions in one line can cause noise in the other lines

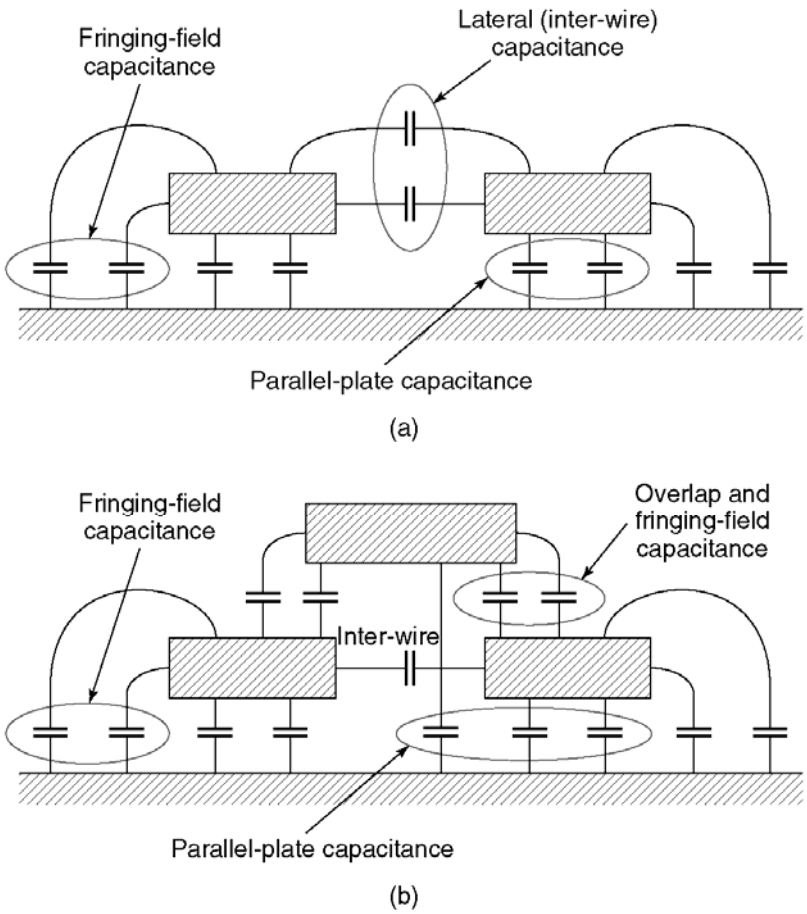


Figure 6.20 Capacitive coupling components, (a) between two parallel lines running on the same level, and (b) between three parallel lines running on two different levels.

Capacitance of an interconnect line

- The capacitance of a line which is coupled with two other lines on both sides
  - If both of the neighboring lines are biased at ground potential
  - The total parasitic capacitance can be more than 20 times as large as the simple parallel-plate capacitance

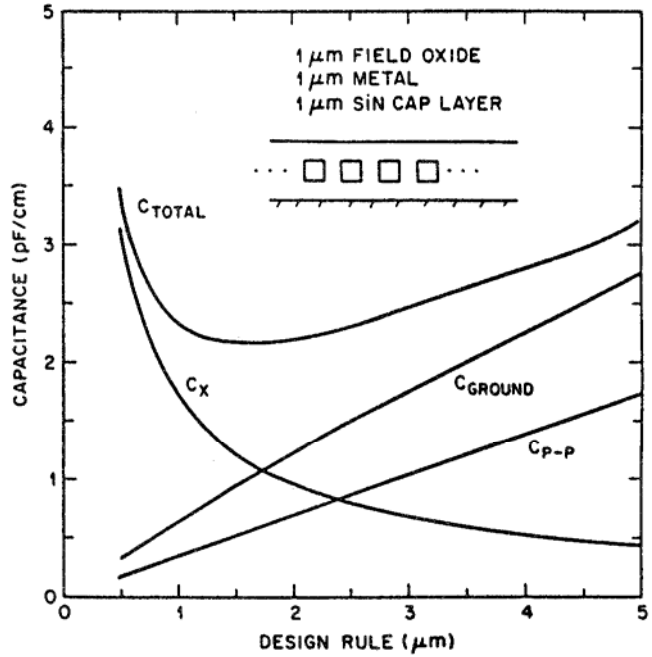


Figure 6.21 Capacitance of an interconnect line which is coupled with two other parallel lines on both sides, as a function of the minimum distance between the lines. C_{TOTAL} indicates the combined capacitance of the line, while C_{GROUND} and C_X indicate the capacitance to the ground plane and the lateral (inter-line) capacitance, respectively. The pure parallel plate capacitance is also shown as a reference.

Capacitance between various layers

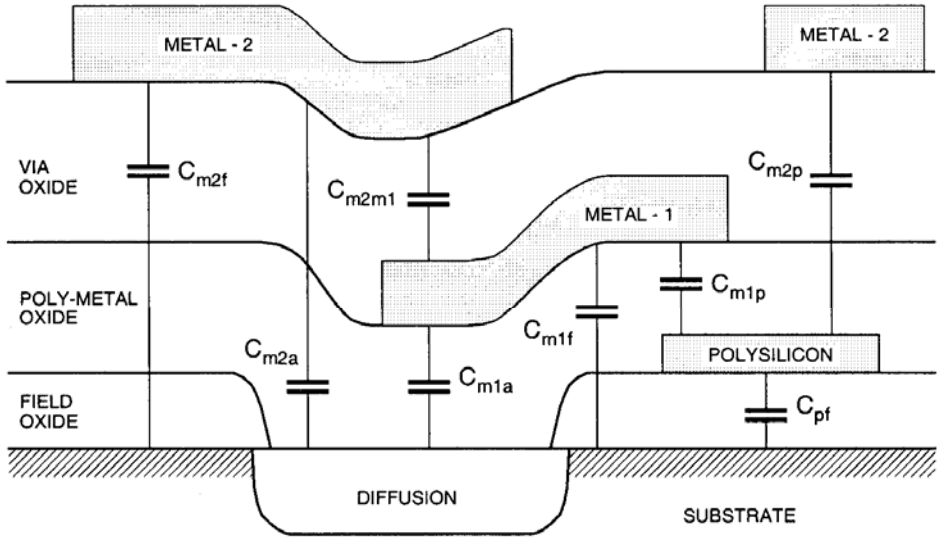


Figure 6.22 Cross-sectional view of a double-metal CMOS structure, showing capacitances between various layers.

Table 6.1 Thickness values of different layers in a typical 0.8 micron CMOS process

Field oxide thickness	0.52 μm	
Gate oxide thickness	16.0 nm	(= 0.016 μm)
Polysilicon thickness	0.35 μm	(minimum width 0.8 μm)
Poly-metal oxide thickness	0.65 μm	
Metal-1 thickness	0.60 μm	(minimum width 1.4 μm)
Via oxide thickness	1.00 μm	
Metal-2 thickness	1.00 μm	(minimum width 1.4 μm)
n^+ junction depth	0.40 μm	
p^+ junction depth	0.40 μm	
n-well junction depth	3.50 μm	

Table 6.2 Parasitic capacitance values between various layers, for a typical double-metal 0.8 micron CMOS technology

Poly over field oxide	C_{pf}	Area	0.066 fF/μm^2
		Perimeter	0.046 fF/μm
Metal-1 over field oxide	C_{m1f}	Area	0.030 fF/μm^2
		Perimeter	0.044 fF/μm
Metal-2 over field oxide	C_{m2f}	Area	0.016 fF/μm^2
		Perimeter	0.042 fF/μm
Metal-1 over Poly	C_{m1p}	Area	0.053 fF/μm^2
		Perimeter	0.051 fF/μm
Metal-2 over Poly	C_{m2p}	Area	0.021 fF/μm^2
		Perimeter	0.045 fF/μm
Metal-2 over Metal-1	C_{m2m1}	Area	0.035 fF/μm^2
		Perimeter	0.051 fF/μm

Interconnect resistance estimation

- The total resistance
  - $R_{wire} = \rho \cdot \frac{l}{w \cdot t} = R_{sheet}\left(\frac{l}{w}\right)$

    R_{sheet} : the sheet resistivity of the line (Ω/square)

    $R_{sheet} = \frac{\rho}{t}$
  - The sheet resistivity
    - Polysilicon: 20-40 Ω/square
    - Silicided ploysilicon: 2-4 Ω/square
    - Aluminum: 0.1 Ω/square
    - Metal-poly, metal-diffusion contact: 20-30 Ω
    - Via resistance: 0.3 Ω
- We can estimate the total parasistic resistance of a wire segment based on its geometry
  - Short distance ⇨negligible
  - Long distance ⇨the total lumped resistance connect in series with the total lumped capacitance

Calculation of interconnect delay- RC delay models

- If the time of flight across the interconnect line is significant shorter than the signal rise/fall times
  - Can be modeled as a lumped RC network
  - Assuming that the capacitance is discharged initially, and assuming that the input signal is a rising step pulse at time t=0
    -

$$V_{out}(t) = V_{DD}\left(1 - \exp\left(-\frac{t}{RC}\right)\right)$$

The rising output voltage reaches the 50% - point at $t = \tau_{PLH}$

$$V_{50\%}(t) = V_{DD}\left(1 - \exp\left(-\frac{\tau_{PLH}}{RC}\right)\right)$$

and the propagation delay for the simple lumped RC network is found as $\tau_{PLH} \approx 0.69RC$

- Unfortunately, this simple lumped RC network provides a very rough approximation
- The accuracy of the simple lumped RC model can be significant improved by
  - Dividing the total resistance into two equal parts
- More accuracy
  - RC ladder network

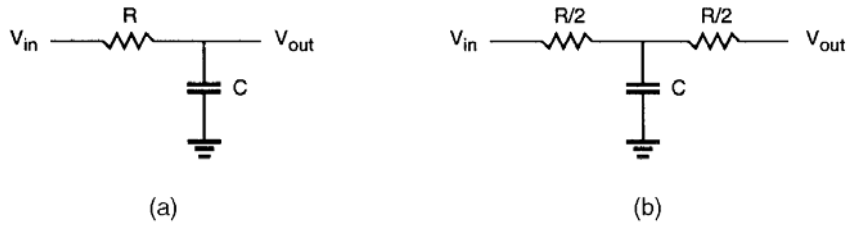


Figure 6.23 (a) Simple lumped *RC* model of an interconnect line, where *R* and *C* represent the total line resistance and capacitance, respectively. (b) The T-model of the same line.

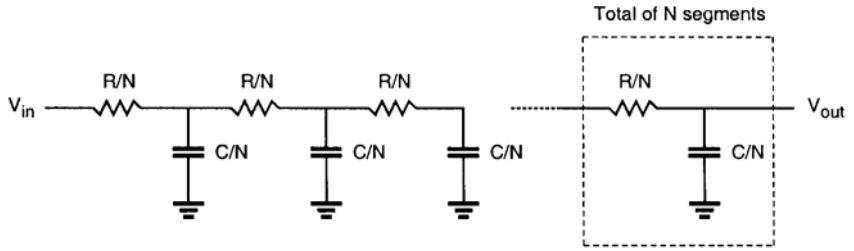


Figure 6.24 Distributed *RC* ladder network model consisting of *N* equal segments.

Calculation of interconnect delay- The Elmore delay

- Consider a general RC tree network
  - There are no resistor loops in this circuit
  - All of the capacitors in an RC tree are connected between a node and a ground
  - There is one input node in the circuit
  - There is a unique resistive path, from the input node to any other node in the circuit
- Path definitions
  - Let P_i denote the unique path from the input node to node i, i=1,2,3..n
  - Let $P_{ij}=P_i \cap P_j$ denote the portion of the path between the input and the node i, which is common to the path between the input and node j

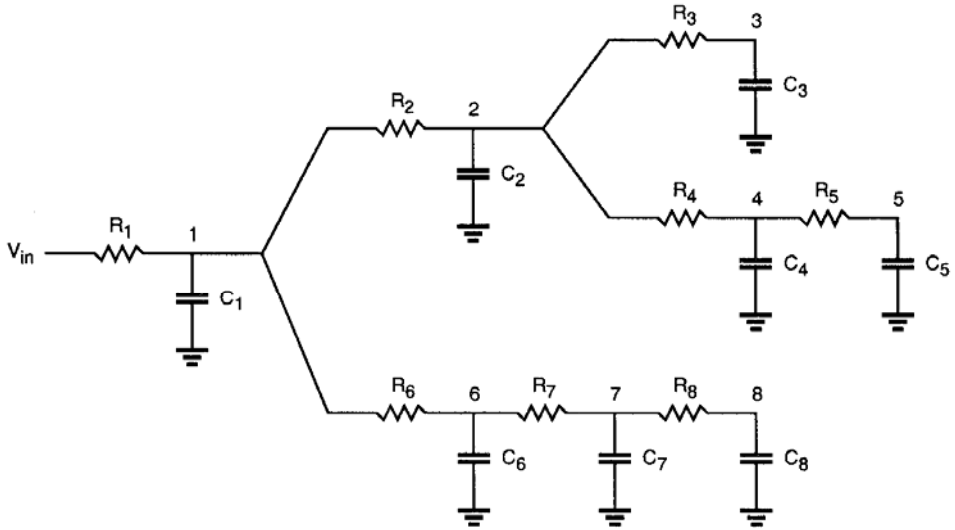


Figure 6.25 A general *RC* tree network consisting of several branches.

Calculation of interconnect delay- The Elmore delay

$$\tau_{D7} = R_1C_1 + R_1C_2 + R_1C_3 + R_1C_4 + R_1C_5 + (R_1 + R_6)C_6 + (R_1 + R_6 + R_7)C_7 + (R_1 + R_6 + R_7)C_8$$

$$\tau_{D5} = R_1C_1 + (R_1 + R_2)C_2 + (R_1 + R_2)C_3 + (R_1 + R_2 + R_4)C_4 + (R_1 + R_2 + R_4 + R_5)C_5 + R_1C_6 + R_1C_7 + R_1C_8$$

A specific case of the general RC tree network $\Rightarrow$ simple RC ladder network

$$\tau_{DN} = \sum_{j=1}^{N} C_j \sum_{k=1}^{j} R_k$$

If assume a uniform RC ladder network, consisting of identical element *(R/N)* and *(C/N)*

$$\tau_{DN} = \sum_{j=1}^{N} \frac{C}{N} \sum_{k=1}^{j} \frac{R}{N} = \left(\frac{C}{N}\right)\left(\frac{R}{N}\right)\frac{N(N+1)}{2} = RC\frac{N+1}{2N}$$

$$\tau_{DN} = \frac{RC}{2} \text{ for } N \to \infty$$

Thus, we see that the propagation delay of a distributed RC line is considerable *smaller* than that of a lumped RC network

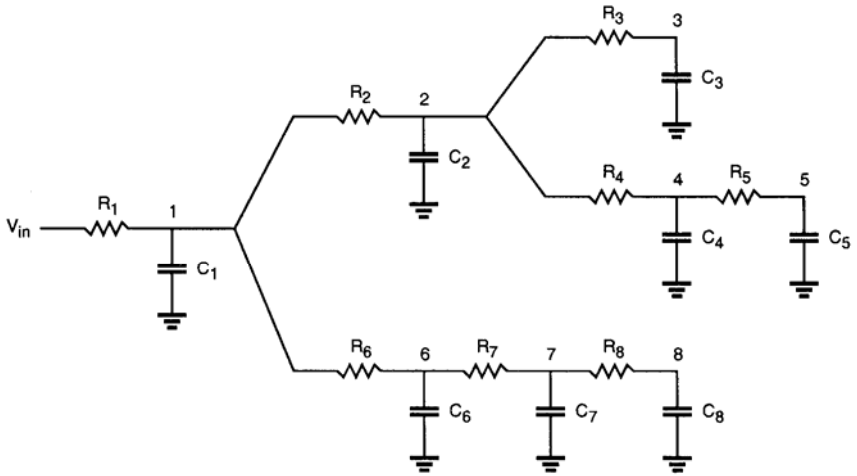


Figure 6.25 A general *RC* tree network consisting of several branches.

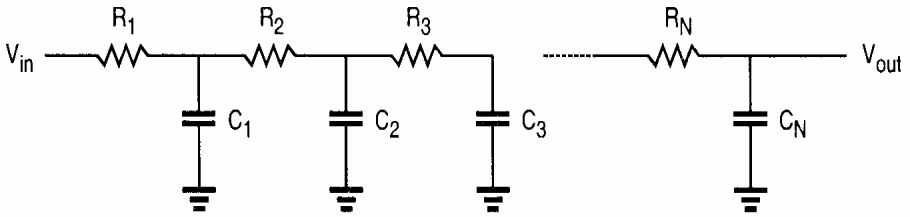


Figure 6.26 Simple *RC* ladder network consisting of one branch.

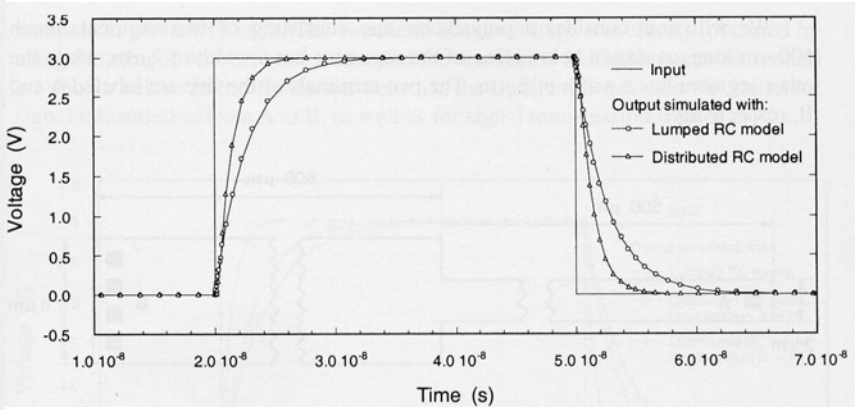
Example 5

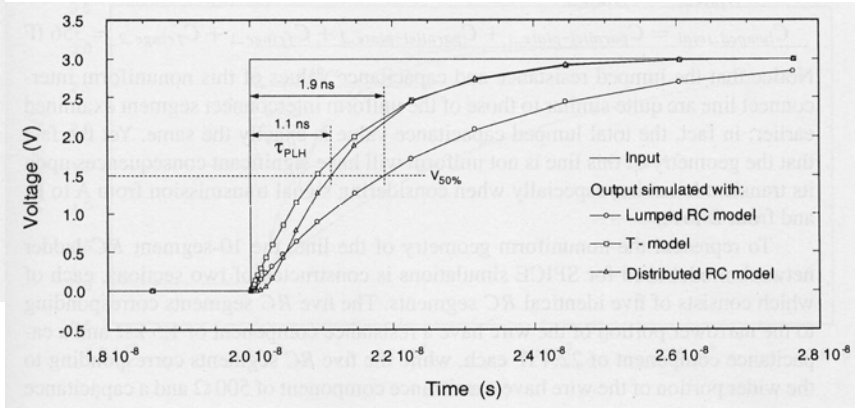
In this example, we will examine the signal propagation delay across a long polysilicon interconnect line, and we will compare various simulation models that can be used to represent the transient behavior of the interconnect. First, consider a uniform polysilicon line with a length of 1000 μm and a width of 4 μm. Assuming a sheet resistance value of 30 Ω/square, the total lumped resistance of the line can be found using (6.55), as

$$R_{lumped} = R_{sheet} \times (\#\text{ of squares})$$
$$= 30\,(\Omega/\text{square}) \times \left(\frac{1000\,\mu\text{m}}{4\,\mu\text{m}}\right) = 7.5\,\text{k}\Omega$$

To calculate the total capacitance associated with this interconnect line, we have to consider both the parallel-plate capacitance component and the fringing-field component. Using the unit capacitance values given in Table 6.2, we obtain

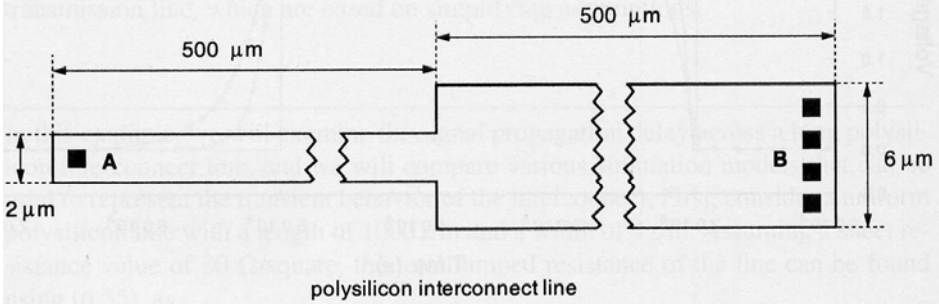
$$C_{parallel\text{-}plate} = (\text{unit area capacitance}) \times (\text{area})$$
$$= 0.066\ \text{fF}/\mu\text{m}^2 \times (1000\,\mu\text{m} \times 4\,\mu\text{m}) = 264\ \text{fF}$$
$$C_{fringe} = (\text{unit length capacitance}) \times (\text{perimeter})$$
$$= 0.046\ \text{fF}/\mu\text{m} \times (1000\,\mu\text{m} + 1000\,\mu\text{m} + 4\,\mu\text{m} + 4\,\mu\text{m}) = 92\ \text{fF}$$
$$C_{lumped\_total} = C_{parallel\text{-}plate} + C_{fringe} = 356\ \text{fF}$$





Example 5

We will now consider a polysilicon line consisting of two segments, each 500 μm long, as shown below. One of the segments has a width of 2 μm, while the other segment has a width of 6 μm. The two terminals of the line are labeled A and B, respectively.



polysilicon interconnect line

The resistance of each segment, as well as the total lumped resistance of the entire interconnect line, can be found using (6.55), as follows.

$$R_{lumped\_1} = 30\,(\Omega/\text{square}) \times \left(\frac{500\,\mu\text{m}}{2\,\mu\text{m}}\right) = 7.5\,\text{k}\Omega$$

$$R_{lumped\_2} = 30\,(\Omega/\text{square}) \times \left(\frac{500\,\mu\text{m}}{6\,\mu\text{m}}\right) = 2.5\,\text{k}\Omega$$

$$R_{lumped\_total} = R_{lumped\_1} + R_{lumped\_2} = 10\,\text{k}\Omega$$

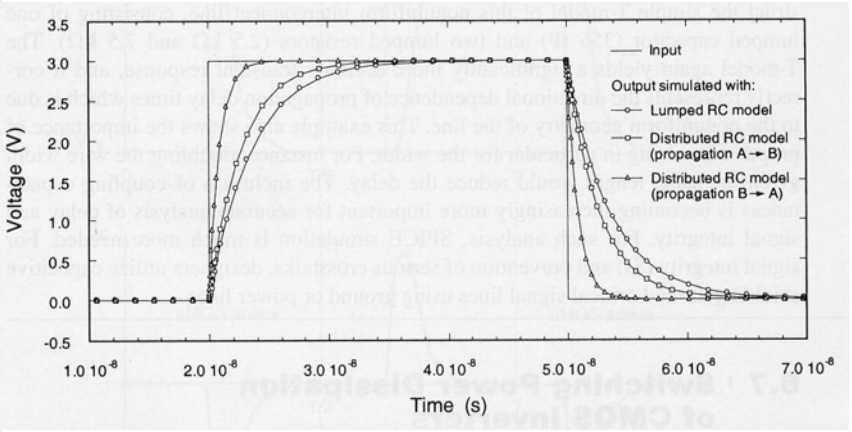
The parasitic capacitances associated with this line are calculated as

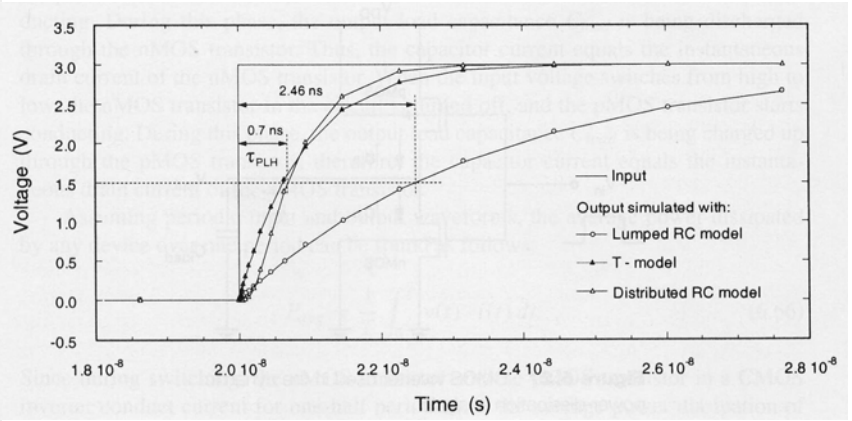
$$C_{parallel\text{-}plate\_1} = 0.066\ \text{fF}/\mu\text{m}^2 \times (500\,\mu\text{m} \times 2\,\mu\text{m}) = 66\ \text{fF}$$

$$C_{parallel\text{-}plate\_2} = 0.066\ \text{fF}/\mu\text{m}^2 \times (500\,\mu\text{m} \times 6\,\mu\text{m}) = 198\ \text{fF}$$

$$C_{fringe\_1} \approx C_{fringe\_2} = 46\ \text{fF}$$

$$C_{lumped\_total} = C_{parallel\text{-}plate\_1} + C_{parallel\text{-}plate\_2} + C_{fringe\_1} + C_{fringe\_2} = 356\ \text{fF}$$





Switching power dissipation of CMOS inverters

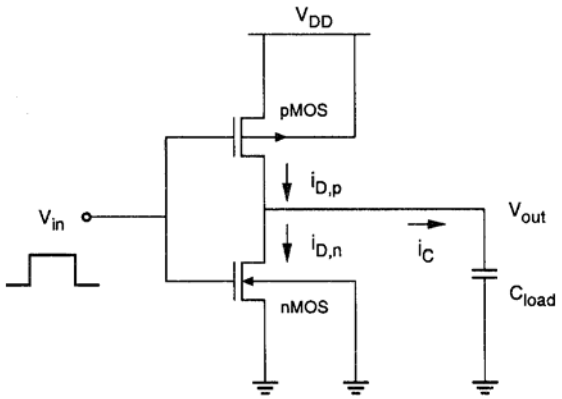


Figure 6.27 CMOS inverter used in the dynamic power-dissipation analysis.

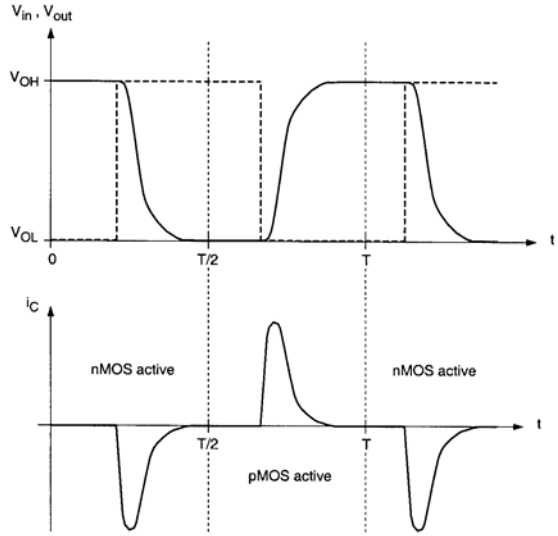


Figure 6.28 Typical input and output voltage waveforms and the capacitor current waveform during switching of the CMOS inverter.

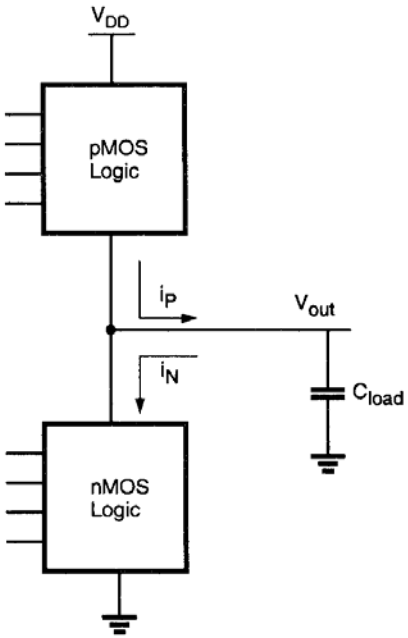


Figure 6.29 Generalized CMOS logic circuit.

$$P_{avg} = \frac{1}{T}\int_0^T v(t)\cdot i(t)dt$$

$$P_{avg} = \frac{1}{T}\left[\int_0^{T/2} V_{out}\left(-C_{load}\frac{dV_{out}}{dt}\right)dt + \int_{T/2}^{T}\left(V_{DD}-V_{out}\right)\left(C_{load}\frac{dV_{out}}{dt}\right)dt\right]$$

$$P_{avg} = \frac{1}{T}\left[\left(-C_{load}\frac{V_{out}^2}{2}\right)\Bigg|_0^{T/2} + \left(V_{DD}\cdot V_{out}\cdot C_{load} - \frac{1}{2}C_{load}V_{out}^2\right)\Bigg|_{T/2}^{T}\right]$$

$$P_{avg} = \frac{1}{T}C_{load}V_{DD}^2$$

$$P_{avg} = C_{load}\cdot V_{DD}^2\cdot f$$

Power meter simulation

- Power meter
  - Estimating the average power dissipation of an arbitrary device or circuit driven by a periodic input, with transient circuit simulation
  - Consisting
    - A linear-controlled current source
    - A capacitor
    - A resistor
  - $$C_y \frac{dV_y}{dt} = \beta i_s - \frac{V_y}{R_y}$$

    The initial condition of the node voltage V_y is aet as $V_y(0) = 0V$

    $$V_y(t) = \frac{\beta}{C_y}\int_0^t \exp\left(-\frac{t-\tau}{R_y C_y}\right) i_{DD}(\tau)\,d\tau$$

    Assuming $R_y C_y >> T$, $V_y(T) \approx \frac{\beta}{C_y}\int_0^T i_{DD}(\tau)\,d\tau$

    If $\beta = V_{DD}\frac{C_y}{T}$ then $V_y(T) = V_{DD}\cdot\frac{1}{T}\int_0^T i_{DD}(\tau)\,d\tau$

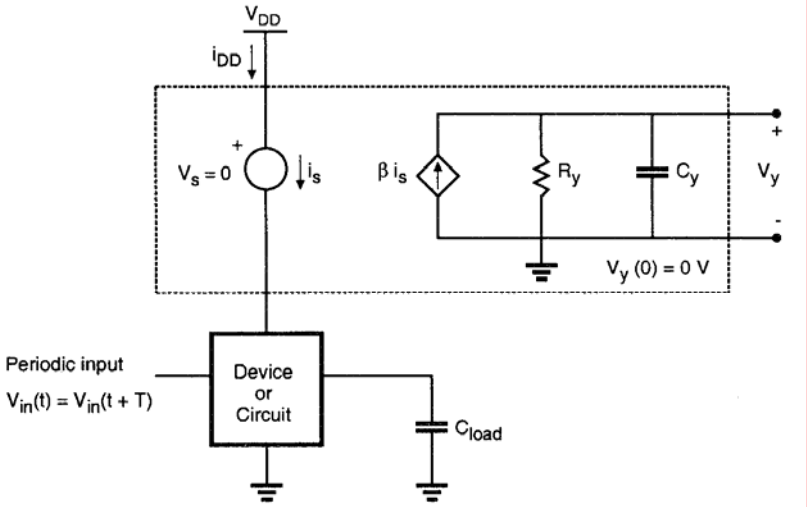


Figure 6.30 The power meter circuit used for the simulation of average dynamic power dissipation of an arbitrary device or circuit.

  - The right-hand side of (6.75) corresponds to the average power drawn from the power supply source over one period
  - The value of the node voltage V_y at t=T gives the average power dissipation

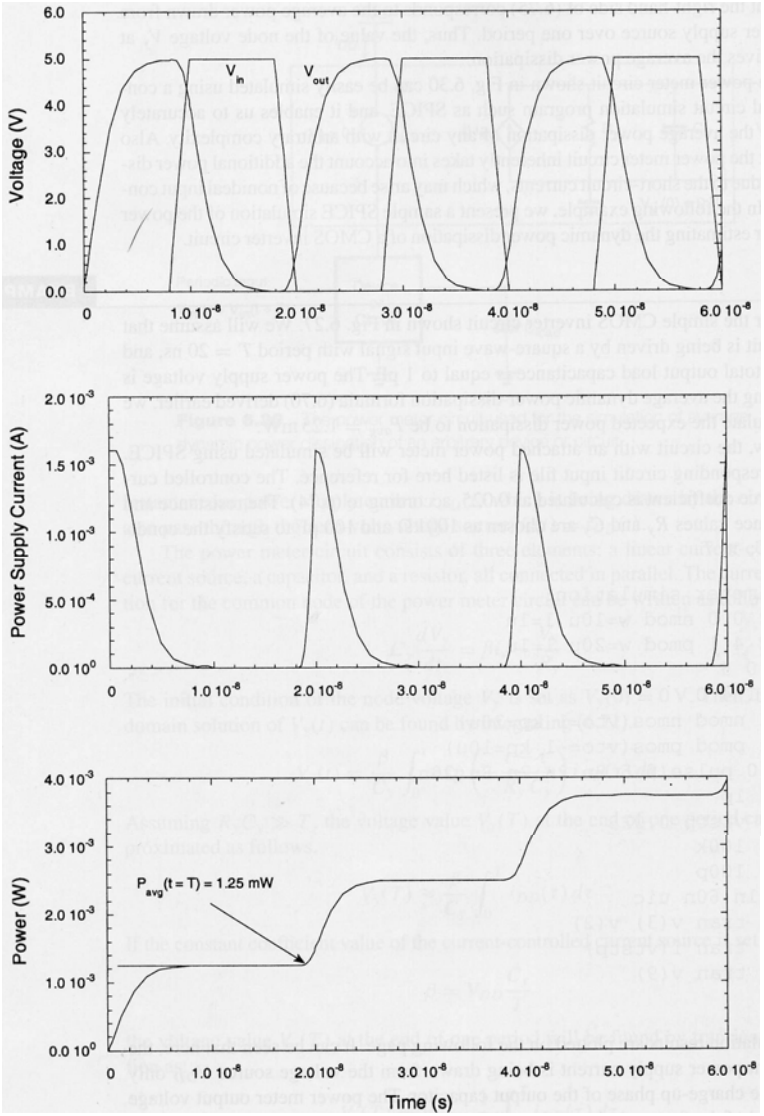
Example 6

Consider the simple CMOS inverter circuit shown in Fig. 6.27. We will assume that the circuit is being driven by a square-wave input signal with period $T = 20$ ns, and that the total output load capacitance is equal to 1 pF. The power supply voltage is 5 V. Using the average dynamic power-dissipation formula (6.70) derived earlier, we can calculate the expected power dissipation to be $P_{avg} = 1.25$ mW.

Now, the circuit with an attached power meter will be simulated using SPICE. The corresponding circuit input file is listed here for reference. The controlled current source coefficient is calculated as 0.025, according to (6.74). The resistance and capacitance values R_y and C_y are chosen as 100 kΩ and 100 pF to satisfy the condition $R_y C_y \gg T$.

```
Power meter simulation:
mn 3 2 0 0 nmod w=10u l=1u
mp 3 2 4 1 pmod w=20u l=1u
vdd 1 0 5
vtstp 1 4 0
.model nmod nmos(vto=1 kp=20u)
.model pmod pmos(vto=-1 kp=10u)
vin 2 0 pulse(0 5 8n 2n 2n 8n 20n)
cl 3 0 1p
fp 0 9 vtstp 0.025
rp 9 0 100k
cp 9 0 100p
.tran 1n 60n uic
.print tran v(3) v(2)
.print tran i(vtstp)
.print tran v(9)
.end
```

The simulation results are plotted on the following page. It can be seen that here, the significant power supply current is being drawn from the voltage source V_{DD} only during the charge-up phase of the output capacitor. The power meter output voltage by the end of the first period corresponds to exactly 1.25 mW, as expected.



Power-delay product

- For measuring the quality and the performance of a CMOS process and gate design
- The average *energy* required for a gate to switch its output voltage from low to high and from high to low
- $PDP=C_{load}V^2_{DD}$ (6.76)
  - Dissipated as heat during switching
  - To keep C_{load} and V_{DD} as low as possible
- $PDP=2P^*_{avg}\,\tau_p$ (6.77)
  - P^*_{avg} is the average switching power dissipation at *maximum* operating frequency
  - τ_p is the average propagation delay
  - The factor of 2, accounting two transitions of the output, from low to high and from high to low
  - This result may *misleading interpretation* that the amount of energy required per switching event is a function of the operating frequency

$$
\begin{aligned}
PDP &= 2\left(C_{load}V_{DD}^2 f_{\max}\right)\tau_p \\
&= 2\left[C_{load}V_{DD}^2\left(\frac{1}{\tau_{PHL}+\tau_{PLH}}\right)\right]\left(\frac{\tau_{PHL}+\tau_{PLH}}{2}\right) \\
&= C_{load}V_{DD}^2
\end{aligned}
$$

Super buffer design (1)

- Super buffer
  - A chain of inverters designed to drive a large capacitive load with minimal signal propagation delay time
- A major objective of super buffer design
  - Given the load capacitance faced by a logic gate, design a scaled chain of N inverters such that the delay time between the logic gate and the load capacitance node is minimized
  - The design task is to determine
    - The number of stages, N
    - The optimal scale factor, α

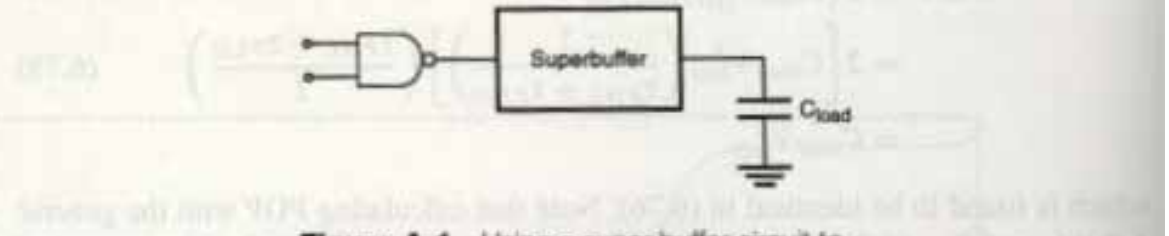


Figure A.1 Using a super buffer circuit to drive a large capacitive load.

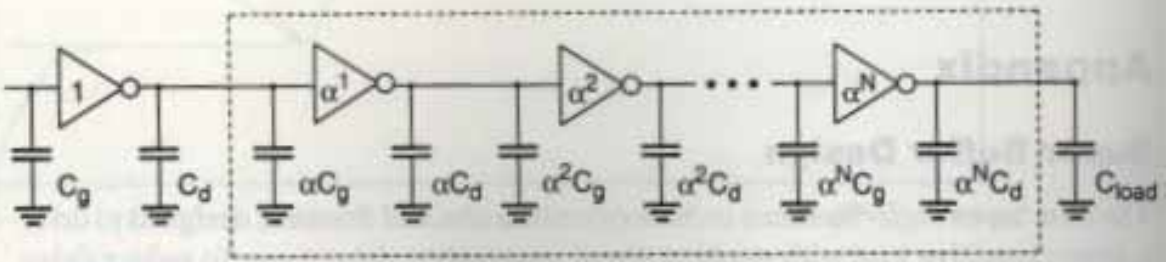


Figure A.2 Scaled super buffer circuit consisting of *N* inverter stages.

Super buffer design (2)

- For the super buffer
  - C_g: the input capacitance of the first stage inverter
  - C_d: the chain capacitance of the first stage inverter
  - The inverters in the chain are scaled up by a factor of α per stage
  - $C_{load} = \alpha^{N+1} C_g$
  - All inverters have identical delay of $\tau_0 (C_d + \alpha C_g)/(C_d + C_g)$
    - τ_0: the per gate delay in the ring oscillator circuit with load capacitance $(C_d + C_g)$

$$\tau_{total} = (N+1)\tau_0 \left(\frac{C_d + \alpha C_g}{C_d + C_g} \right)$$

$$From \;\; C_{load} = \alpha^{N+1} C_g \Rightarrow (N+1) = \frac{\ln\left(\frac{C_{load}}{C_g} \right)}{\ln \alpha}$$

$$\tau_{total} \doteq \frac{\ln\left(\frac{C_{load}}{C_g} \right)}{\ln \alpha} \tau_0 \left(\frac{C_d + \alpha C_g}{C_d + C_g} \right)$$

$$\frac{\partial \tau_{total}}{\partial \alpha} = \tau_0 \ln\left(\frac{C_{total}}{C_g} \right) \left[-\frac{\frac{1}{\alpha}}{(\ln \alpha)^2} \left(\frac{C_d + \alpha C_g}{C_d + C_g} \right) + \frac{1}{\ln \alpha} \left(\frac{C_g}{C_d + C_g} \right) \right] = 0$$

the optimal scale factor $\alpha(\ln \alpha - 1) = \frac{C_d}{C_g}$

A special case of the above equation occurs when the drain capacitance is neglected; i.e., $C_d = 0$. In that case, the optimal scale factor becomes the natural number $e = 2.718$, However, in reality the drain parasitics cannot be ignored