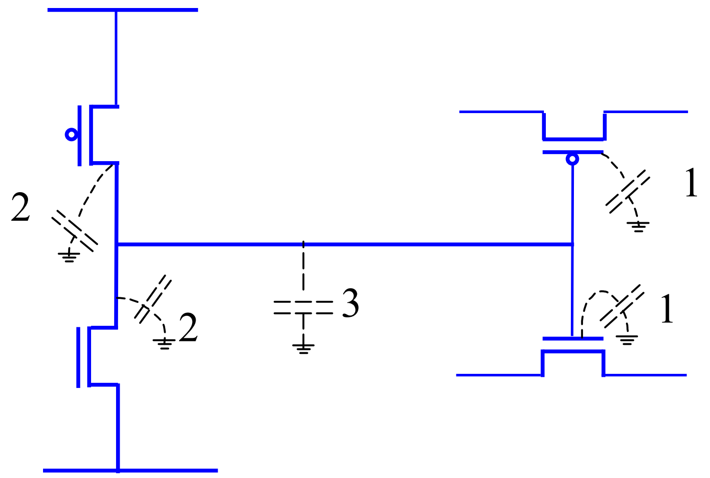
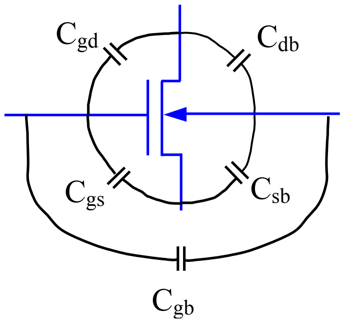
The times in these transitions are clearly determined by load, capacitive or resistive, driven by an inverter. Hence it is important to determine this load. Primary timing comes from capacitive load and hence we will only consider this.

Load capacitance estimation



When a CMOS gate drive another CMOS gate, the driver "sees" a capacitive load. The load capacitance consists of (1) gate capacitance of the load, (2) diffusion capacitance of the driver, and (3) the routing capacitance.

Intrinsic Capacitances of MOS



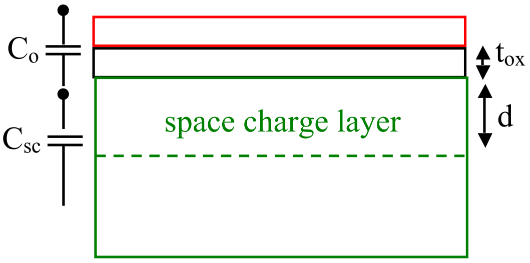
(Identify these capacitances from p. 4 MOSFET structure)

Gate Capacitance

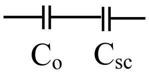
When transistor is "OFF", there is no channel, there are two dielectric regions below gate – oxide (gives capacitance C_o) and space charge region (gives capacitance C_{SC}) in the substrate as shown. $C_{gs} = C_{gd} = 0$. ε_{SiO2} and ε_{Si} are dielectric constants of oxide and silicon, respectively. A is area of the channel region (=WL), W being the MOSFET width and L MOSFET length.

$C_{gb} = C_o$ in series with $C_{sc.}$

$$C_o = \frac{\varepsilon_{sio_2}}{t_{ox}} A$$



For $0 < V_{gs} < V_{th}$, a depletion layer of depth d is formed in the substrate.

C_{gb} = 

$C_{sc} = \frac{\varepsilon_{si}}{d} A$ and decreases with increasing V_{gs}.

Hence C_{gb} decreases with increasing V_{gs} as d increases. However, this is generally simplified so that

$C_{gb} = C_o$.

In the linear region, a continuous channel from source to drain is formed. The gate to channel capacitance is $\frac{\varepsilon_{sio_2} A}{t_{ox}}$.

Let $\frac{1}{2}$ of this be C_{gs} and the other half be C_{gd} .

Hence ,

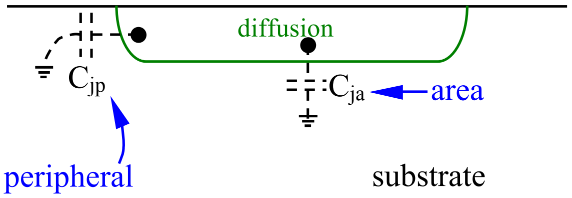
$$C_{gs} = C_{gd} = \frac{\varepsilon_{sio_2} A}{2t_{ox}}$$

In the saturation region , the drain region of the channel is pinched-off . Here $C_{gd} = 0$ and

$$C_{gs} = \frac{2}{3} \cdot \frac{\varepsilon_{sio_2} A}{t_{ox}}.$$

Parameter	Capacitance		
	Off	Linear	Saturation
C_{gb}	$\frac{\varepsilon_{sio_2} A}{t_{ox}}$	0	0
C_{gs}	0	$\frac{\varepsilon_{sio_2} A}{2t_{ox}}$	$\frac{2\varepsilon_{sio_2} A}{3t_{ox}}$
C_{gd}	0	$\frac{\varepsilon_{sio_2} A}{2t_{ox}}$	0

Diffusion capacitance



Peripheral capacitance is determined by the perimeter of Source or Drain.

Zero bias capacitance

Parameter	n-diffusion	p-diffusion
C_{ja0}	$0.1\,\mathrm{fF}/\mu\mathrm{m}^2$	$0.1\,\mathrm{fF}/\mu\mathrm{m}^2$
C_{jp0}	$0.9\,\mathrm{fF}/\mu\mathrm{m}$	$0.8\,\mathrm{fF}/\mu\mathrm{m}$

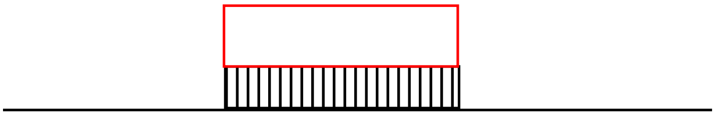
$$C_j = area * C_{ja0}\left(1-\frac{V_j}{\phi}\right)^{-ma} + perimeter * C_{jp0}\left(1-\frac{V_j}{\phi}\right)^{-mp}$$

V_j = junction voltage. (+ve for forward bias. -ve for reverse bias (MOS case).)

ϕ = built-in junction potential ≈ 0.6-0.9 V

ma, mp = 0.3 ~ 0.5

Routing capacitance

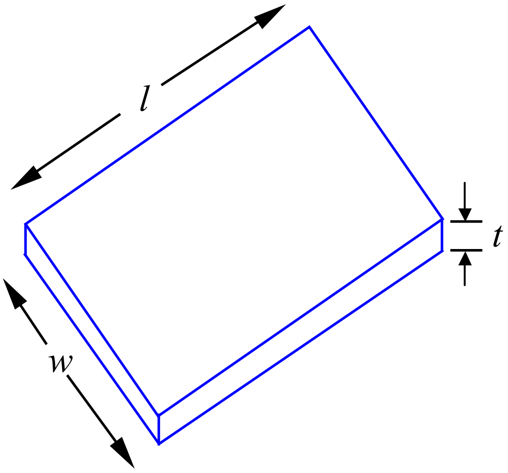


Representative CMOS Values

Capacitance	
Gate (Cox)	$1.918 \pm 0.075\,\mathrm{fF}/\mu\mathrm{m}^2$
Poly2 to Substrate	$0.084 \pm 0.006\,\mathrm{fF}/\mu\mathrm{m}^2$
Poly1 to Poly2	$0.650 \pm 0.050\,\mathrm{fF}/\mu\mathrm{m}^2$
Metal1 to Substrate	$0.041 \pm 0.004\,\mathrm{fF}/\mu\mathrm{m}^2$
Metal1 to Diffusion	$0.080 \pm 0.006\,\mathrm{fF}/\mu\mathrm{m}^2$
Metal1 to poly2	$0.066 \pm 0.004\,\mathrm{fF}/\mu\mathrm{m}^2$
Metal2 to Substrate	$0.019 \pm 0.004\,\mathrm{fF}/\mu\mathrm{m}^2$
Metal2 to Diffusion	$0.029 \pm 0.004\,\mathrm{fF}/\mu\mathrm{m}^2$
Metal2 to Poly2	$0.042 \pm 0.004\,\mathrm{fF}/\mu\mathrm{m}^2$
Metal2 to Metal1	$0.042 \pm 0.004\,\mathrm{fF}/\mu\mathrm{m}^2$

Another important parameter to extract is the resistance of various regions. We will not use it in simple calculations but it is used in accurate simulation.

Sheet resistance



Resistance of a uniform slab

$$R = \left(\frac{\rho}{t}\right)\left(\frac{l}{w}\right), \qquad \rho = \text{resistivity}$$

$$R = \underset{\uparrow}{\rho_s}\left(\frac{l}{w}\right)$$

sheet resistivity

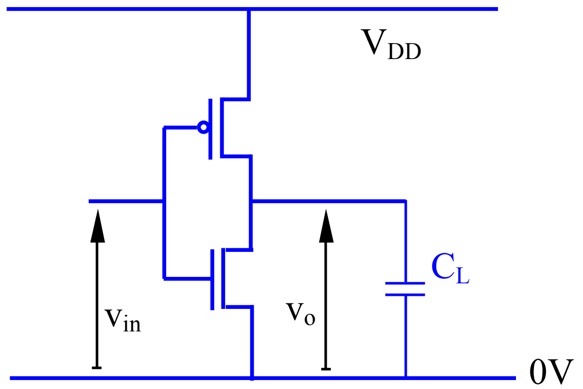
	sheet resistance $\Omega/sq.$		
material	min.	typical	max.
n well	1150	1300	1450
n^+	47	57	67
p^+	85	100	115
poly 1	17.5	22.5	27.5
poly 2	15	20	25
metal 1	0.066	0.072	0.078
metal 2	0.033	0.036	0.039

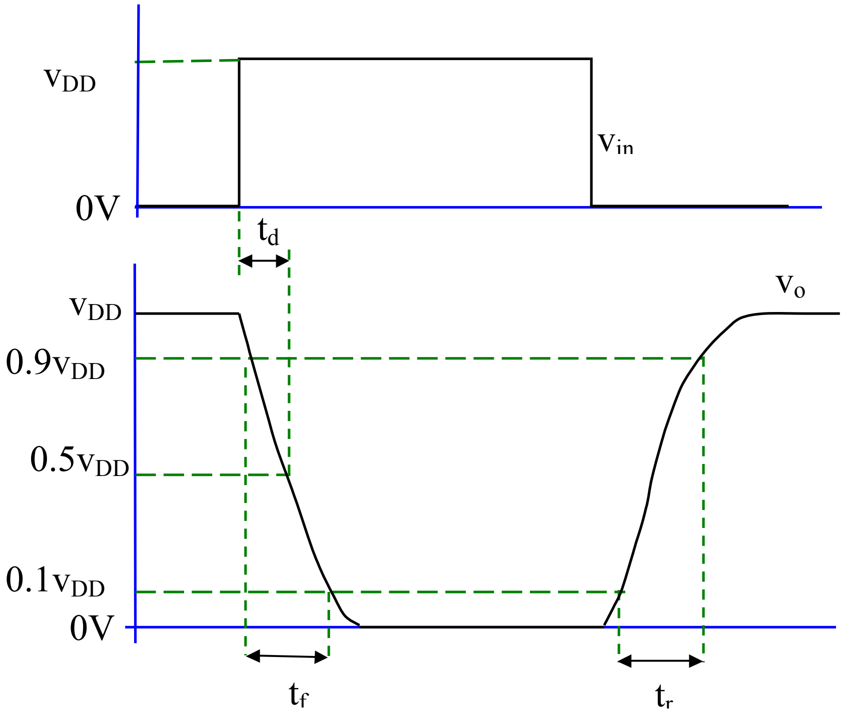
Sheet resistivity of transistor channel in the linear region

$$\rho_s = \left[\frac{\mu\varepsilon}{t_{ox}}\left(V_{gs} - V_t\right)\right]^{-1}$$

Timing Calculation

We will simplify this assuming a fixed load capacitance.





t_d : delay time

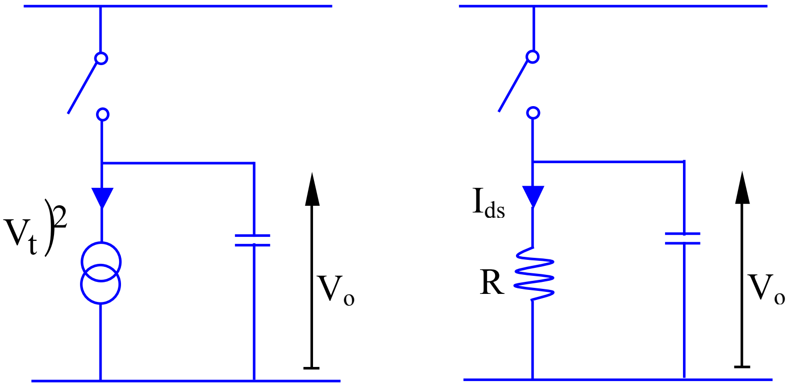
t_f : fall time

t_r : rise time

Fall time (time for V_o to fall from $0.9V_{DD}$ to $0.1V_{DD}$) , t_f

$$I_{ds} = \frac{\beta_n}{2}\left(V_{gs} - V_t\right)^2$$

$$V_{gs} = V_{DD}$$



saturation linear

For $V_o = 0.9V_{DD}$ NMOS in Saturation , $V_o > V_{DD} - V_t$:

V_o drops from $0.9V_{DD}$ to $V_{DD} - V_t$

$$t_{f1} = \frac{C_L\left(0.9V_{DD} - \left(V_{DD} - V_t\right)\right)}{I_{ds}} = \frac{C_L\left(V_t - 0.1V_{DD}\right)}{I_{ds}}$$

$$= \frac{2C_L\left(V_t - 0.1V_{DD}\right)}{\beta_n\left(V_{DD} - V_t\right)^2}$$

For $V_t = 0.2\ V_{DD}$

$$t_{f1} = \frac{0.313C_L}{\beta_n V_{DD}}$$

For $V_o < V_{DD} - V_t$, NMOS in Linear region : V_o drops

from $V_{DD} - V_t$ to $0.1V_{DD}$

$$I_{ds} = \beta_n\left[\left(V_{gs} - V_t\right)V_{ds} - \frac{V_{ds}^2}{2}\right]$$

$$= \beta_n\left[\left(V_{DD} - V_t\right)V_o - \frac{V_o^2}{2}\right]$$

$$t_{f2} = \int\frac{C_L dV_o}{-I_{ds}}$$

$$= \frac{C_L}{\beta_n}\int_{V_{DD}-V_t}^{0.1V_{DD}} \frac{-dV_o}{\left(V_{DD} - V_t - \frac{V_o}{2}\right)V_o}$$

$$= \frac{C_L}{\beta_n\left(V_{DD} - V_t\right)}\ln\left\{\frac{19V_{DD} - 20V_t}{V_{DD}}\right\}$$

For $V_t = 0.2\ V_{DD}$

$$t_{f2} = \frac{C_L}{0.8\beta_n V_{DD}}\ln(15) = \frac{3.39C_L}{\beta_n V_{DD}}$$

$\therefore$ fall time, $t_f = t_{f1} + t_{f2} = \dfrac{3.7C_L}{\beta_n V_{DD}}$

If $\beta_p = \beta_n$, **rise time = fall time,** same condition that is required for right switching voltage. This is thus normally adopted.

Delay time , t_d

Saturation : V_o drops from V_{DD} to V_{DD} - V_t

$$t_{d1} = \frac{C_L\left(V_{DD} - \left(V_{DD} - V_t\right)\right)}{I_{ds}} = \frac{C_L V_t}{I_{ds}}$$

$$= \frac{2C_L V_t}{\beta_n \left(V_{DD} - V_t\right)^2}$$

For $V_t = 0.2\ V_{DD}$, $t_{d1} = \dfrac{0.625C_L}{\beta_n V_{DD}}$

Linear : V_o drops from V_{DD} - V_t to 0.5 V_{DD}

$$t_{d2} = \frac{C_L}{\beta_n}\int_{V_{DD}-V_t}^{0.5V_{DD}} \frac{-dV_o}{\left(V_{DD} - V_t - \frac{V_D}{2}\right)V_o}$$

$$= \frac{C_L}{\beta_n\left(V_{DD} - V_t\right)}\ln\left(\frac{3V_{DD} - 4V_t}{V_{DD}}\right)$$

For $V_t = 0.2\ V_{DD}$, $t_{d2} = \dfrac{C_L \ln(2.2)}{0.8\beta_n V_{DD}} = \dfrac{0.986C_L}{\beta_n V_{DD}}$

$\therefore$ delay time $= t_{d1} + t_{d2} = \dfrac{1.61C_L}{\beta_n V_{DD}}$

Technology Innovation? How to reduce these delays and get faster circuits?

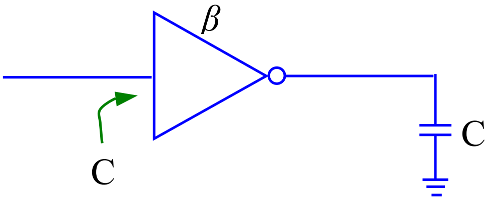
Hence for a fixed load, W of the devices can be increased to achieve the right required delay. Such fixed capacitances are normally dominated by the interconnect capacitance.

However, if device W is increased, the area of the channel (WL) and area/perimeter of source/drain also automatically increase. Hence the previous stage now has to drive a larger intrinsic gate capacitance load. This needs to be tackled right so that the delay can be minimised. The issues are explained in the following section.

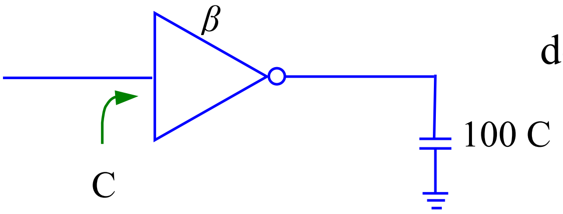
Driving intrinsic large capacitance load

Let input capacitance seen at the gates of inverter sized for equal rise and fall time *i.e.*

both transistors have $\beta_n = \beta_p = \beta$ where β is a given constant.



$$\text{delay} = \tau_o = K\,C/\beta, \text{ where K is a constant.}$$

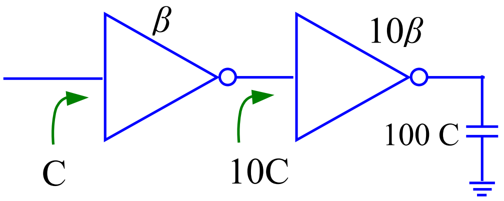


$$\text{delay} = K\,(100C)/\beta = 100\,\tau_o$$

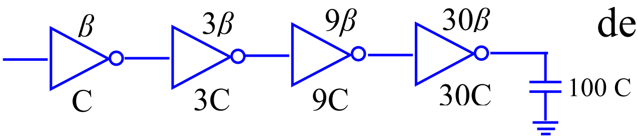
One may think that by increasing $\beta_n = \beta_p$ by a factor of 100 to 100 β will reduce the delay to τ_o. However, the capacitance seen at the input now becomes 100C as the gate area also increases by the factor 100 and delay at the input stage will now rise. Hence this basically transfers delay issue to the input. Hence a better strategy will be to drive using a chain of progressively growing inverters to obtain minimum delay as shown below.

Here, for the last stage, the load is 100C, $\beta_n = \beta_p = 10\beta$.

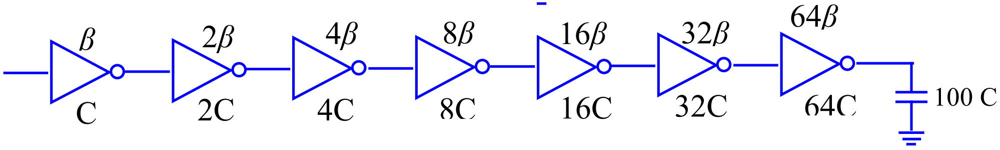
Hence the delay of the last stage = $K*100C/(10\beta) = 10KC/\beta = 10\ \tau_o$. For the first stage, load is 10C and $\beta_n = \beta_p = \beta$. Hence the delay of the first stage = $K*(10C)/\beta = 10KC/\beta = 10\ \tau_o$. The delay of every stage is similarly calculated.



delay $= 10\ \tau_o + 10\ \tau_o$
$= 20\ \tau_o$

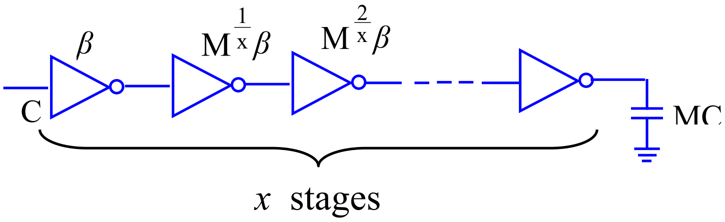


delay = $(3 + 3 + 3.3 + 3.3)\tau_o$
$= 12.6\ \tau_o$



delay = $(2 + 2 + 2 + 2 + 2 + 2 + 1.6)\ \tau_o = 13.6\ \tau_o$

Now look at a general problem with load = MC.



Total number of stages is x and each inverter size grows by a factor $M^{\frac{1}{x}}$. Hence load for the first stage will be $M^{\frac{1}{x}}$C and it will progressively grow by a factor $M^{\frac{1}{x}}$. Hence for any arbitrary n^{th} stage, the load capacitance will be $CM^{\frac{n}{x}}$ and size factor will be $M^{\frac{n-1}{x}}\beta$. Hence the delay of the n^{th} stage is

$$K\frac{Load\,Capacitance}{Size\,Factor}=K\frac{CM^{\frac{n}{x}}}{\beta M^{\frac{n-1}{x}}}=K\frac{CM^{\frac{1}{x}}}{\beta}=M^{\frac{1}{x}}\tau_0$$

as the delay of τ_o is for the ratio C/β. As delay of every stage is the same and there are x stages, the total delay $=\tau=\tau_o\,x\,M^{\frac{1}{x}}$

Find x which minimizes τ.

$$\log_e\tau=\log_e x+\frac{1}{x}\log_e M+\log_e\tau_o$$

$$\frac{d\log_e\tau}{dx}=\frac{1}{x}-\frac{\log_e M}{x^2}=0$$

$$\therefore \quad x = \log_e M$$

$$\therefore\ M^{\frac{1}{x}} = e$$

$\therefore$ for optimum switching speed , each inverter should be e times larger than the previous one.

Scaling of MOS Devices

Influence of first-order scaling on MOS device characteristics

	PARAMETERS	SCALING FACTOR
DEVICE PARAMETERS	Length; L	$1/\alpha$
	Width: W	$1/\alpha$
	Gate oxide thickness; t_{ox}	$1/\alpha$
	Junction depth; X_j	$1/\alpha$
	Substrate doping; $N_{a\ (\text{or } b)}$	α
	Supply voltage; V_{DD}	$1/\alpha$
RESULTANT INFLUENCE	Electric field across gate oxide; E	1
	Depletion layer thickness; d	$1/\alpha$
	Parasitic capacitance; WL/t_{ox}	$1/\alpha$
	Gate delay; (VC/I)	$1/\alpha$
	DC power dissipation; P_s	$1/\alpha^2$
	Dynamic power dissipation; P_d	$1/\alpha^2$
	Power-speed product	$1/\alpha^3$
	Gate area	$1/\alpha^2$
	Power density; (VI/A)	1
	Current density; (I/A)	α
	Transconductance; g_m	1

In summary, CMOS technology offers tremendous power management advantage.

The switching characteristics and symmetric rise/fall delay dictate that the width of PMOS device is twice that of NMOS device in an inverter. Expressions for switching voltage and time delay can be derived for any simple MOSFET I-V model.