LESSON PLAN

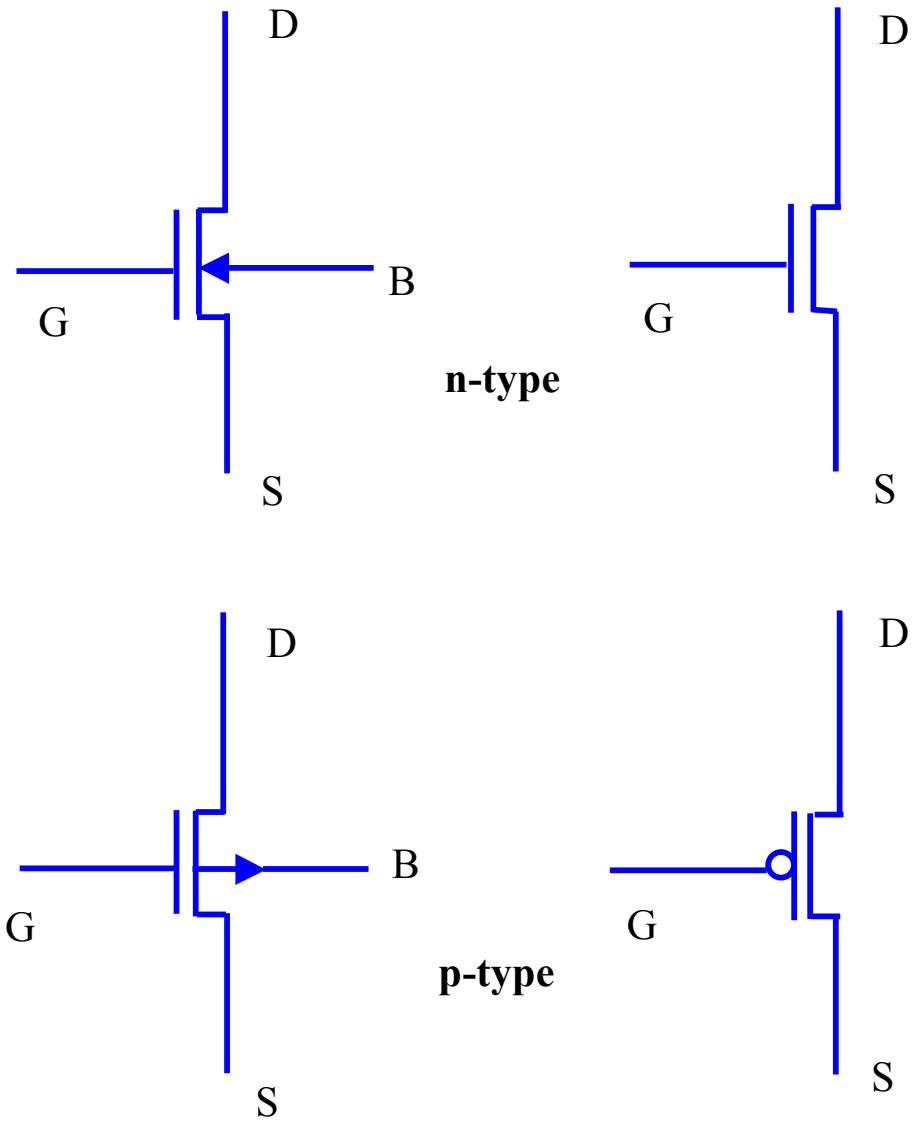
You have learnt by now how to synthesize digital circuits with the help of standard cell library and hardware description language. Now we move on to learn how these cells are designed, how there functionality and speed can be implemented and how they are realized on Silicon.

The first part on CMOS inverter technology will introduce basic device and describes switching characteristics of inverter. Delay calculations and optimization then follow. CMOS-Layout part concentrated on processes and inverter cross sections so that ideas on mask and what they do are clear. Main focus is on designing masks to implement logic circuits. Both one metal layer and two metal layer versions are included. Finally, module concludes with computational elements, memory and system design example.

There will be at least one question on physical layout or stick diagram or a combination. Other questions will be distributed over all other topics.

CMOS Technology

In CMOS technologies, both n-type (n-channel) and p-type (p-channel) transistor are used. S denotes Source, D the Drain, G the Gate and B the body or substrate terminals.



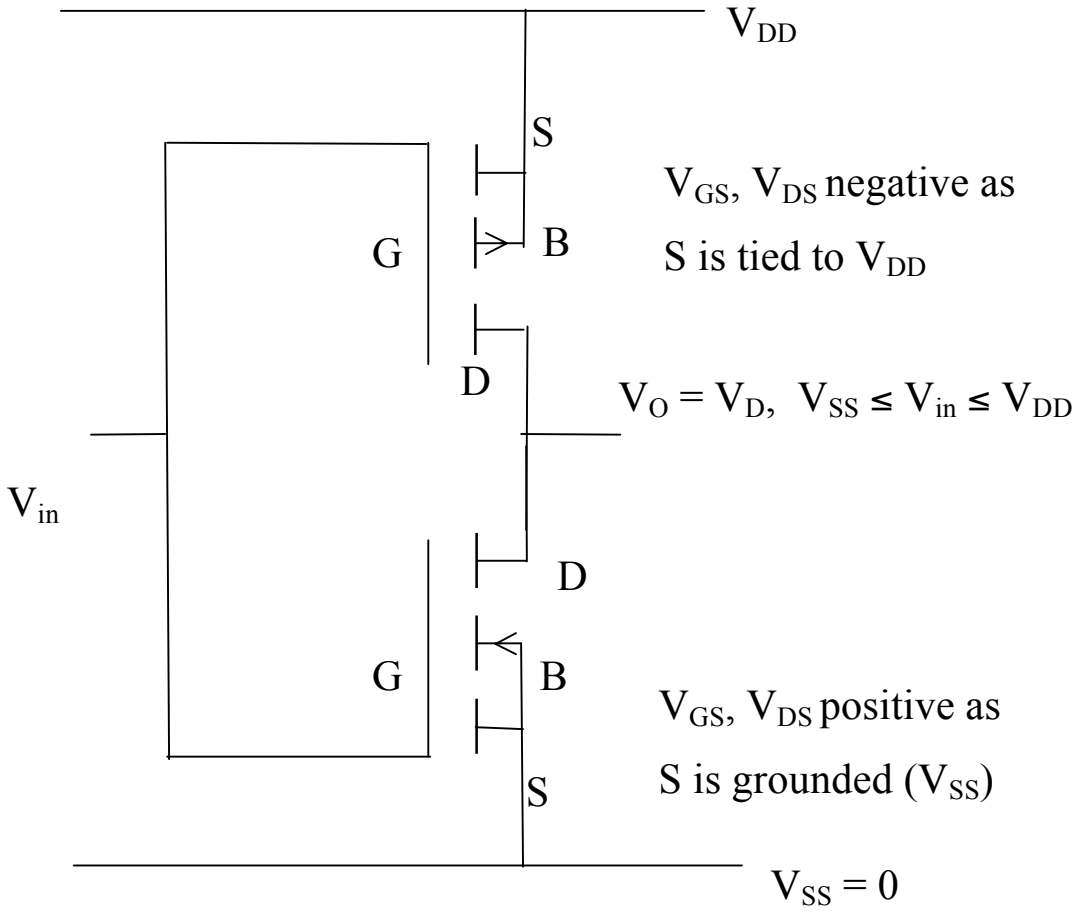
n-type transistor : substrate is p-doped (acceptor type) and is connected to -ve supply. Usually zero.

p-type transistor : substrate is n-doped (donor type) connected to +ve supply. Usually power supply voltage.

For simplicity, as substrate bias is usually fixed, removal of substrate (body) contact leads to the compact notation on the right.

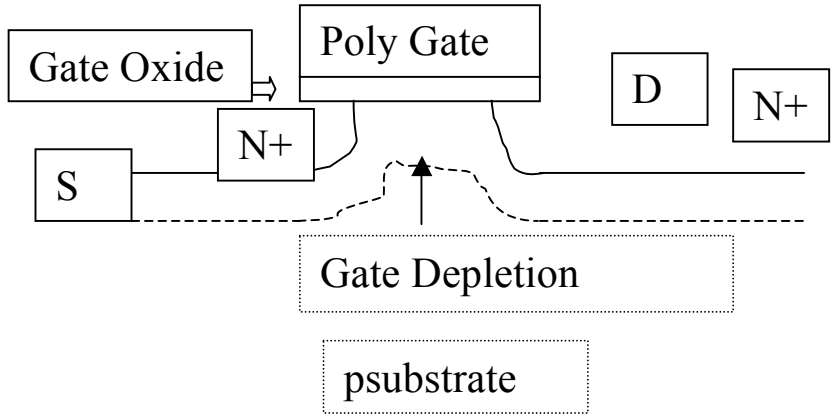
The following circuit shows another way to represent 4 terminal MOS device (transistor).

Basic simplest circuit block in CMOS technology is an inverter that uses both devices. The schematic of CMOS inverter is shown below.



Here, top one is a p channel device and lower one is an n channel device. Source, body, gate and drain terminals are all labeled.

Basic n-channel MOS Device



- Device Description

- Thin gate oxide always desirable for better gate control. Typically 7.5 nm for 0.25 micron gate length technology. As low as 1.2 nm for technologies below 0.1 micron.

- Polysilicon gate for self alignment that helps to achieve a better process integration (clarified later).

- Heavily doped source drain regions for low resistance.

- Low resistance metal connections for source, drain, body (same as substrate or well or tub different notations).

- If V_{GS} is negative, holes in the substrate will be attracted to Si-SiO_2 interface. Accumulation.

- As V_{GS} increases, the interface comes out of accumulation and the gate depletion (or space charge

region) shown by dashed line under the gate becomes deeper.

- If V_{GS} increases further, eventually the device enters the regime of strong inversion where there is a large electron concentration (opposite to the substrate carriers holes) below the oxide. The gate depletion now stops growing. The threshold voltage of MOS device is defined as the V_{GS} that leads to strong inversion. After this V_{GS}, the device conducts with the help of the electrons for positive V_{DS}.

- P-channel device has opposite relative voltages, has P+ (heavily p doped) source-drains, and n-type substrate. Both devices are used in CMOS technology which is our focus. P-channel device normally carries less current for similar dimension compared to n-channel due to lower mobility of holes.

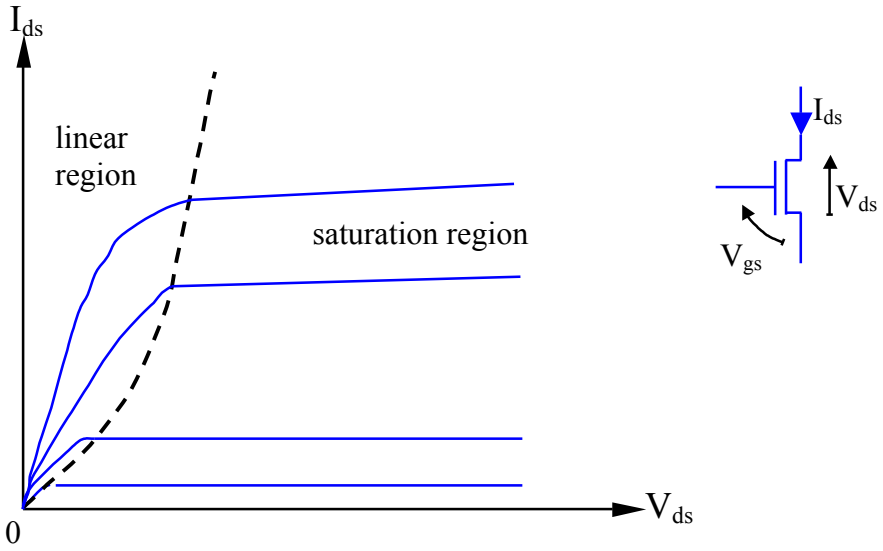
- For designers, it is important to know how the drain current I_D changes when V_{GS} and V_{DS} change. Under simplified assumptions, analytic expressions for this current can be derived which we will use for design.

- I-V plots and analytic expressions will now be discussed followed by cell design and timing calculations with parasitic (unwanted capacitors and resistors which cannot be removed) estimation.

- Much later, we will move on to study how to fabricate these devices so that you have basic knowledge on what determines device limitations and design rules

and how to logically develop layout of a circuit to implement it in Silicon.

MOS transistor characteristics



Simple Model

cut-off region : $V_{gs} - V_t \leq 0$

V_t: threshold voltage

$I_{ds} = 0$

(V_t also denoted as V_{TN} for n-channel MOSFET and V_{TP} for p-channel MOSFET)

linear region : $0 < V_{ds} \leq V_{gs} - V_t$

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

saturation region : $0 < V_{gs} - V_t \leq V_{ds}$

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

$$\beta = \frac{\mu\varepsilon}{t_{ox}}\left(\frac{W}{L}\right)$$

μ : effective surface mobility of charge carrier.

ε : permittivity of gate insulator.

t_{ox} : thickness of gate insulator.

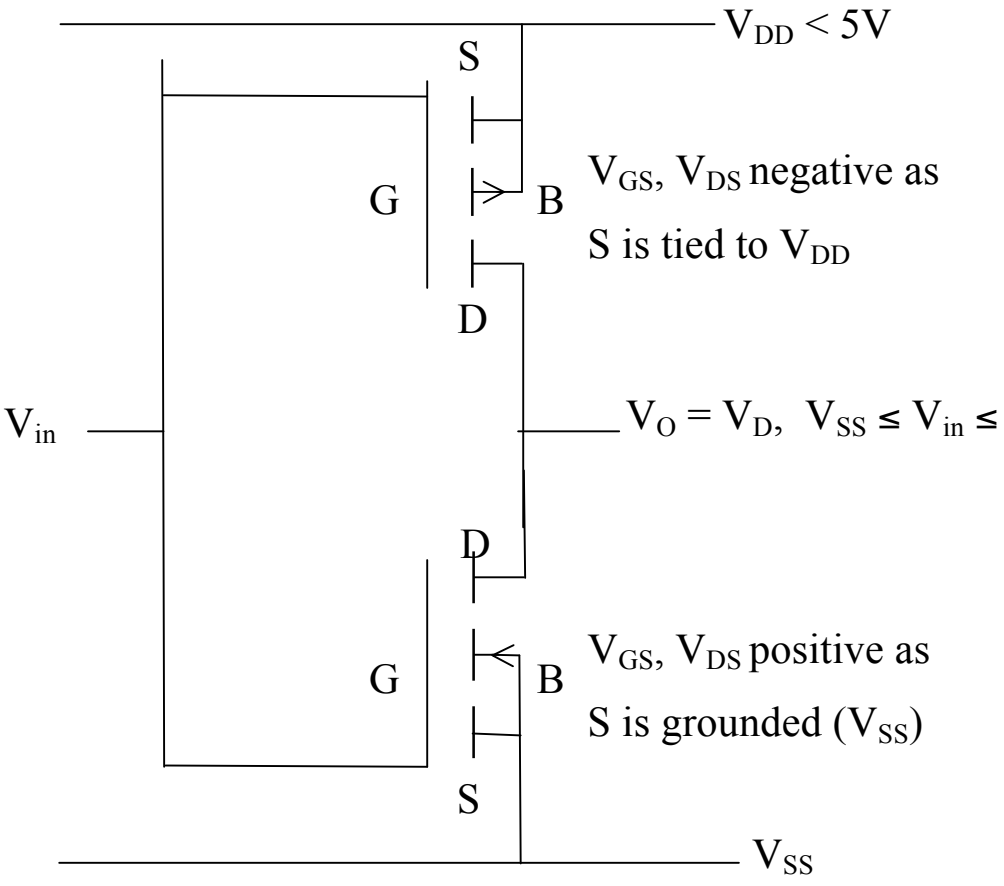
W : channel width.

L : channel length.

Technology Innovation

Basic Operation of an Inverter

The schematic of CMOS inverter is repeated below.



Here, top one is a p channel device and lower one is an n channel device. Source, body, gate and drain terminals are all labeled.

Qualitatively, it is easy to see how the inverter functions. If $V_{in} = 0$, then n-channel device is off. Hence there is no current through the n-channel device and consequently through the p-channel device (by KCL). For the p-channel device, $V_{GS} = V_G - V_S = -V_{DD}$. Hence, it is turned on. The only consistent voltage, where $V_{GS} = -V_{DD}$ and p-channel device current is zero, is V_{DS} (p-channel) = 0 i.e. $V_O = V_{DD}$. Thus, if $V_{in} = 0$, $V_O = V_{DD}$.

On the other hand, if $V_{in} = V_{DD}$, then $V_{GS} = 0$ for the p-channel device. Since $V_{GS} \leq V_{TP}$ is a necessity for it to conduct, the device is off. Hence, there is no current through the n-channel device which is turned on by $V_{in} = V_{DD} \geq V_{TN}$. Again, for this to be consistent,

$$V_{DS} \text{ (n-channel)} = 0 \text{ or } V_O = 0.$$

Hence if $V_{in} = V_{DD}$, $V_O = 0$.

Thus it is clear that the circuit here performs a job of an inverter with $V_{in} = 0$ or $V_{in} = V_{DD}$. When input switches

from say 0 to V_{DD}, output switches from V_{DD} to 0. The devices are in different regions of operation during the switch. The switching of the inverter now will be discussed.

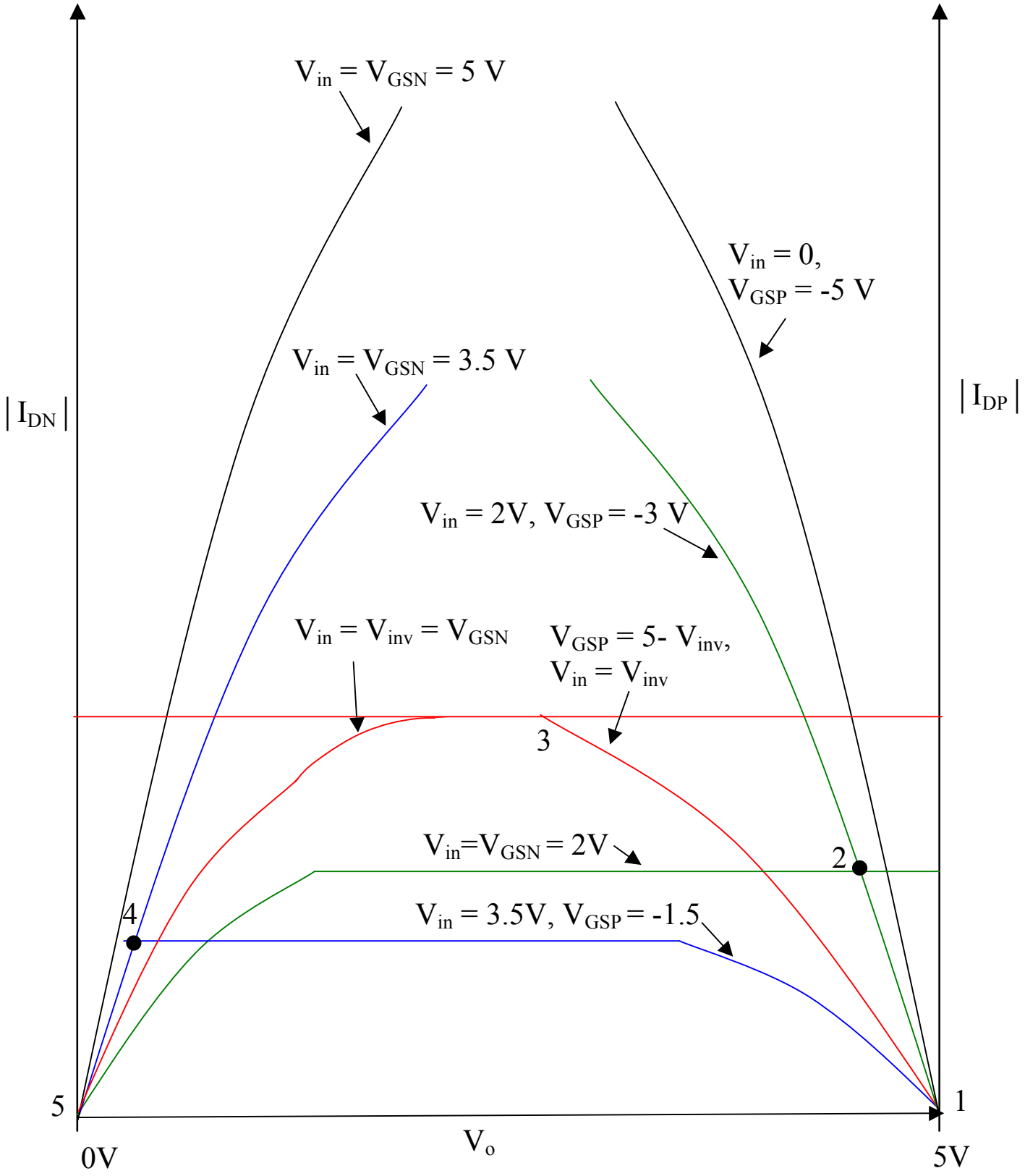
Switching Characteristics of the Inverter

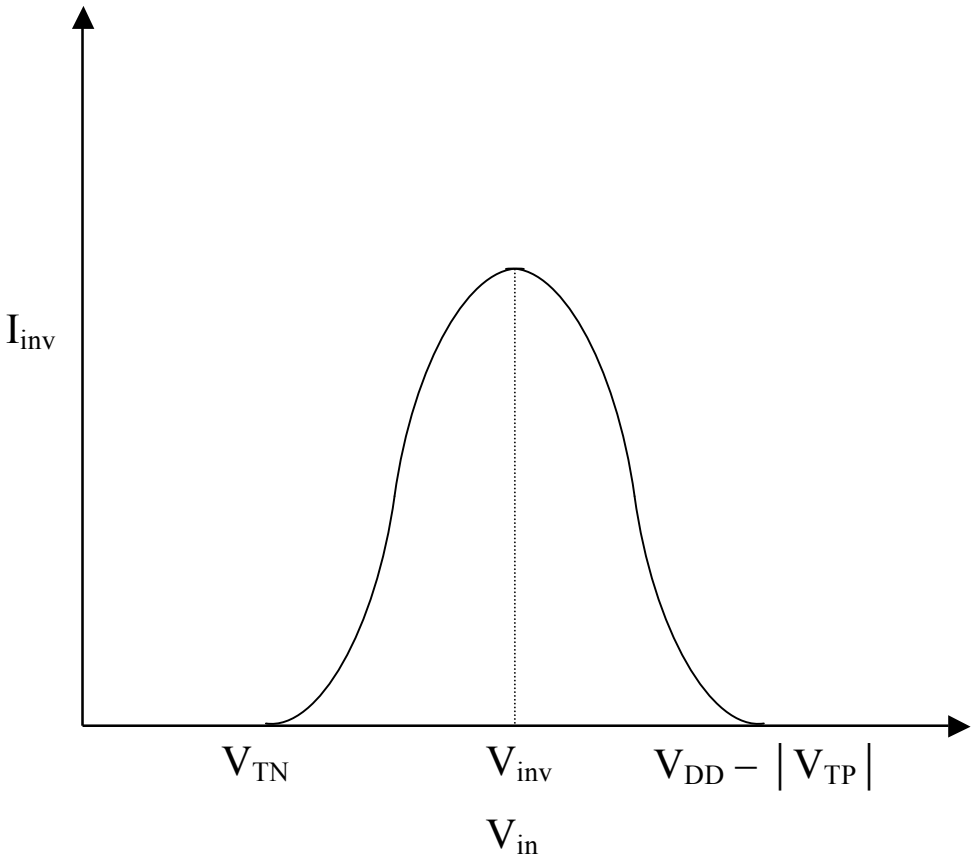
As V_{in} changes, these devices go through five different combinations of regions of operation. These combinations are indicated below.

1.	$0 \le V_{in} \le V_{TN}$	n-channel off p-channel on
2.	$V_{TN} < V_{in} < V_{inv}$	p-channel linear n-channel saturation
3.	$V_{in} = V_{inv}$	Both in saturation
4.	$V_{inv} < V_{in} < V_{DD} - V_{TP}$	n-channel linear p-channel saturation
5.	$V_{DD} - V_{TP} \le V_{in} \le V_{DD}$	p-channel off n-channel on

Where V_{inv} is called as switching voltage. The characteristics are derived based on the fact that both MOSFETs in the inverter here carry the same amount of current.

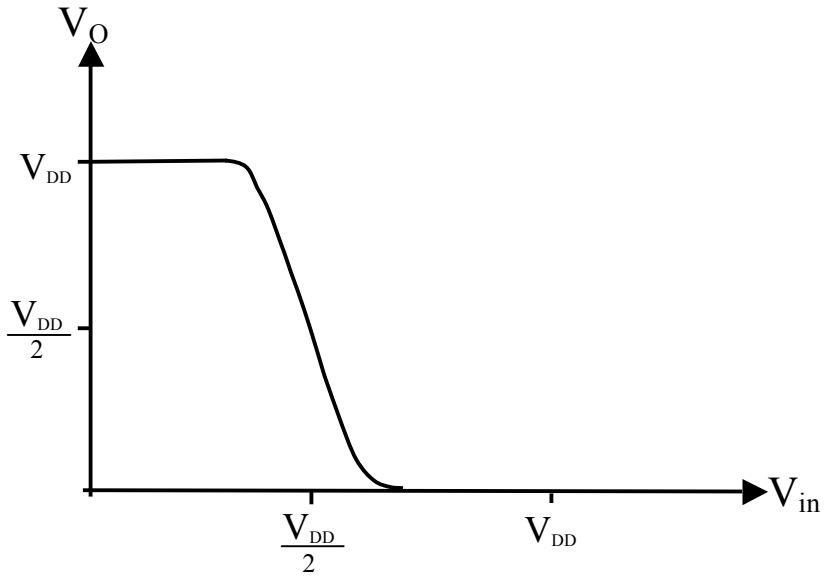
Plot of Switching Current Characteristics





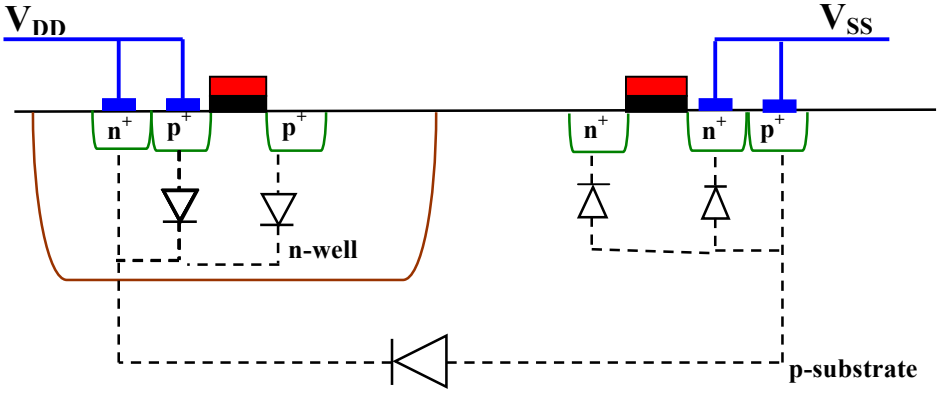
One can notice that for a step V_{in}, the current only flows for a short time. If input is changing in a finite time, the above current will flow through the inverter, which peaks at V_{inv}. This inverter will consume lower amount of power compared to circuits that continuously conduct such as bipolar or NMOS, as there is no current for some voltages. If there is no switching (DC), there is no power consumption, in principle. In reality, there will be some power consumption due to sub-threshold currents, reverse biased diode currents and clocks, which are always running.

Voltage Response (Most change occurs near V_{inv})



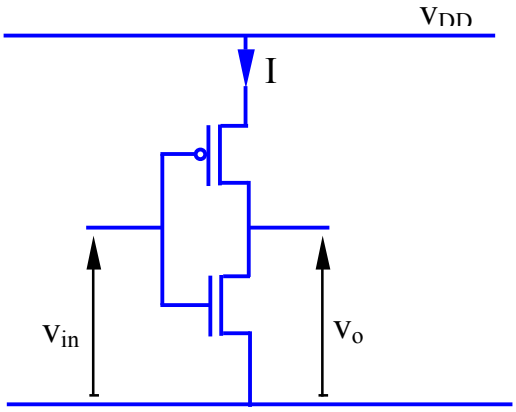
Static Power Consumption

Static current



Static current flows through bottom parasitic diode. In addition, one more diode from p+ or n+ will also leak depending on the output voltage. Also, one of the MOS devices, which is off, will conduct leakage current that will lead to static power consumption.

Switching voltage of inverter



For n-transistor in saturation,

$$I = \frac{\beta_n}{2}(V_{in} - V_t)^2$$

For p-transistor in saturation,

$$I = \frac{\beta_p}{2}(V_{DD} - V_{in} - V_t)^2$$

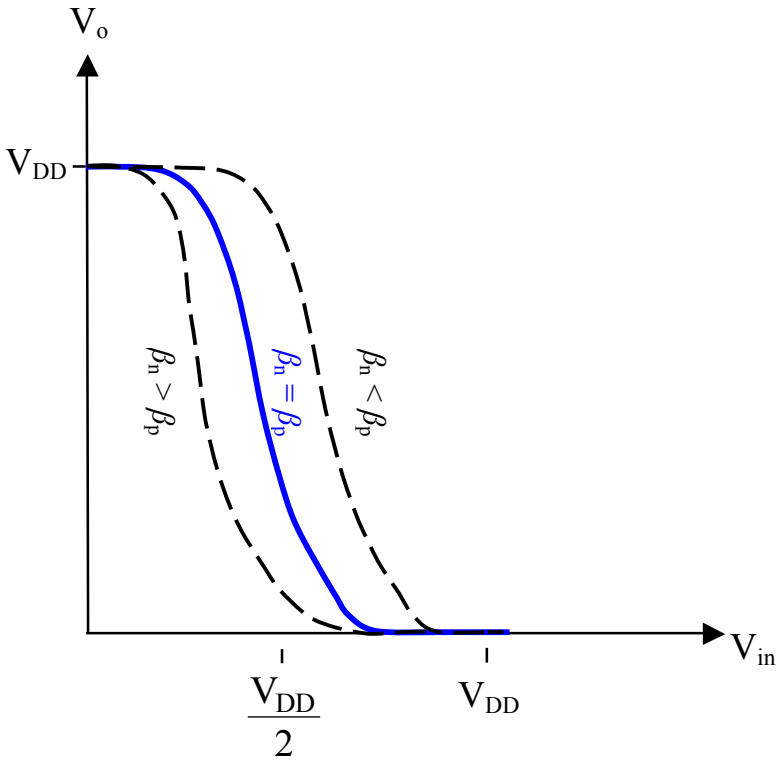
(Assume threshold voltages of n- and p- transistors have the same magnitudes.)

At switching voltage, the saturation current of n- and p-channel devices are equal.

For switching at $V_{in} = V_{DD}/2$,

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

i.e. $\beta_n = \beta_p$

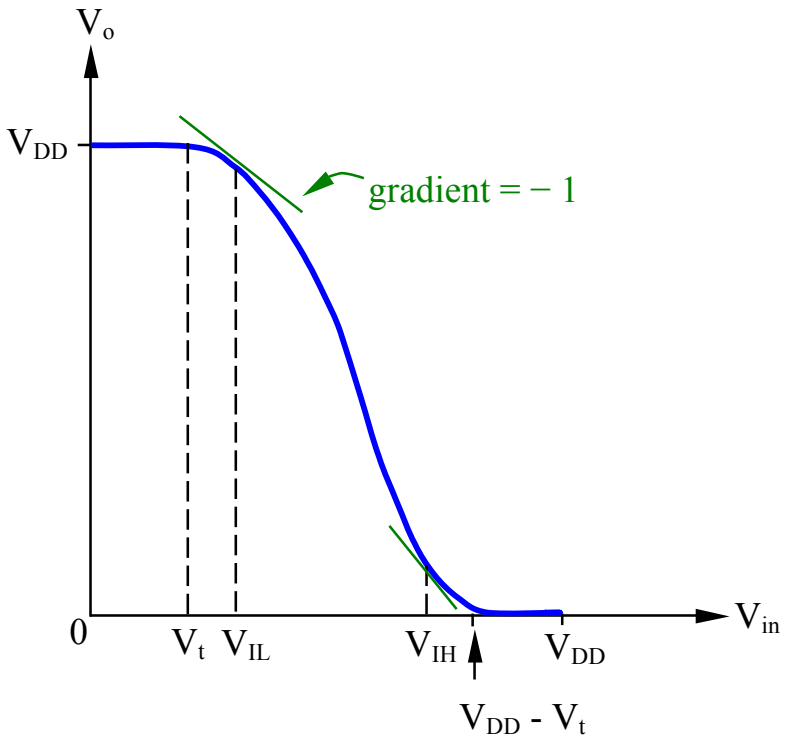


$$\beta_n = \beta_p \Rightarrow \frac{\mu_n \varepsilon}{t_{ox}}\frac{W_n}{L_n} = \frac{\mu_p \varepsilon}{t_{ox}}\frac{W_p}{L_p}$$

i.e. $\frac{W_p}{W_n} = \frac{\mu_n}{\mu_p}$ (for $L_p = L_n$ which is normally the case)

i.e. $\frac{W_p}{W_n} \approx 2$

Noise margin



V_{IH} = minimum HIGH input voltage

V_{IL} = maximum LOW input voltage

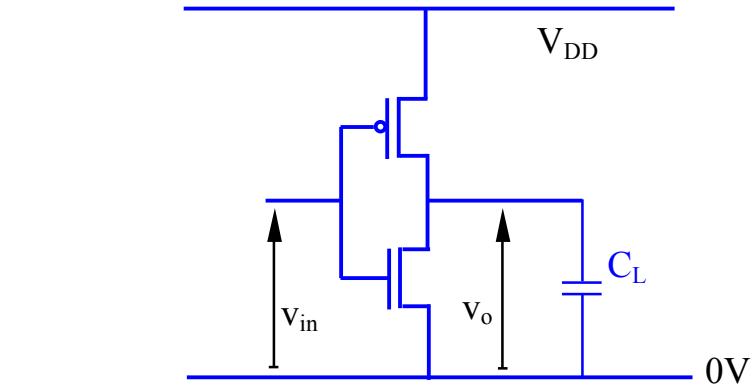
Noise margin (low) = V_{IL}

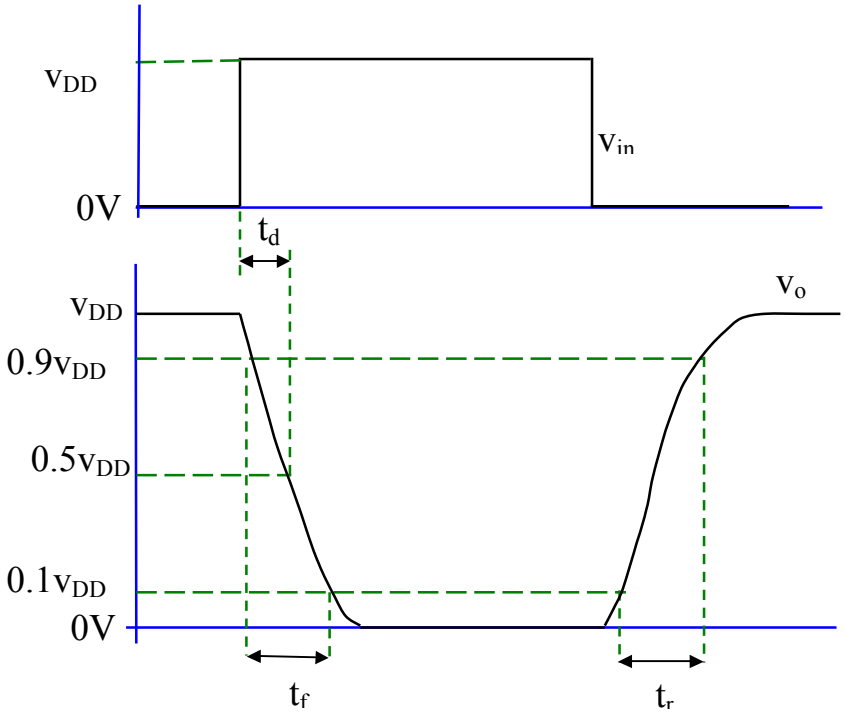
Noise margin (high) = V_{DD} - V_{IH}

Loaded CMOS inverter

However, above conditions are all ideal. Normally, CMOS inverter will have parasitic resistances and capacitances attached to the output node. Hence principle

of operation of inverter alters a bit. The loaded CMOS inverter is shown below.





t_d : delay time

t_f : fall time

t_r : rise time

If input was at low voltage, the output is high and the capacitor C_L holds the charge to maintain this high voltage.

If input now switches instantaneously to high, the capacitor voltage does not change instantaneously as this requires infinite current. Instead, at the switching time, both the gate voltage and the drain voltage are a high for the NMOS transistor which turns on. The NMOS device current will discharge the capacitor to zero after a delay time. This is the reason NMOS device is called as a pull down device. Once the output voltage reaches zero, NMOS transistor stops conducting as $V_{DS} = 0$. The output stays at zero volts after this delay and there is no current in the circuit. Exactly opposite cycle occurs when input becomes low and PMOS device charges the capacitor to raise output voltage to the power supply voltage. Similarly, PMOS device is called as a pull up

device. The various time delays and the waveforms are depicted above. Now we can estimate dynamic power dissipation quite easily during one switching cycle.

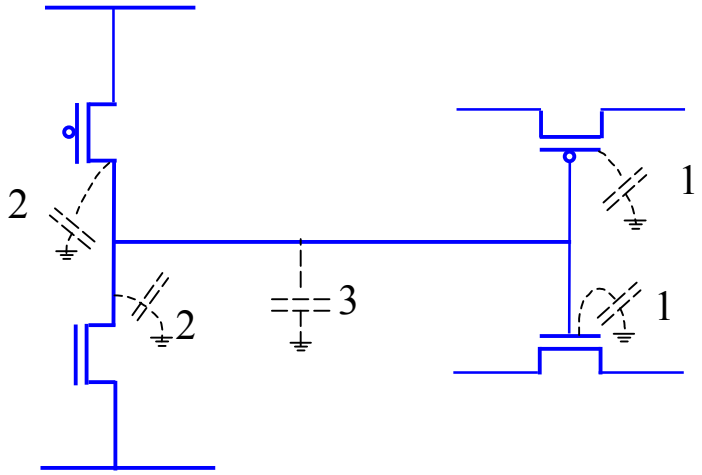
Dynamic power

(1) Charging and discharging of load capacitance determines the dynamic power consumption. If the load capacitance is C, the charge stored is CV_{DD}. ∴ energy supplied by V_{DD} in one full switching cycle is CV_{DD}^2. If the switching frequency is f, the power dissipated (energy drawn from power source in unit time) is fCV_{DD}^2.

How we can minimize dynamic power consumption?

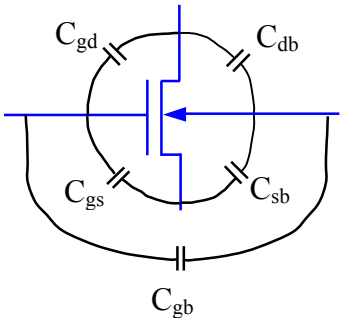
The times in these transitions are clearly determined by load, capacitive or resistive, driven by an inverter. Hence it is important to determine this load. Primary timing comes from capacitive load and hence we will only consider this.

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



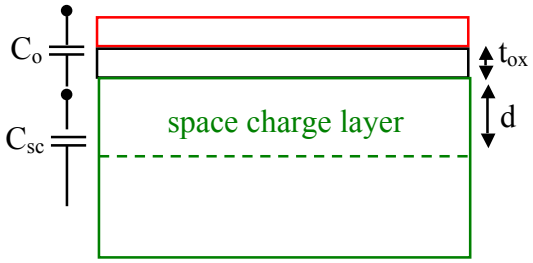
(Identify these capacitances from p. 4 MOSFET structure)

Gate Capacitance

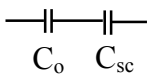
When transistor is "OFF", there is no channel, there are two dielectric regions below gate – oxide (gives capacitance C_o) and space charge region (gives capacitance C_{SC}) in the substrate as shown. $C_{gs} = C_{gd} = 0$. ε_{SiO2} and ε_{Si} are dielectric constants of oxide and silicon, respectively.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

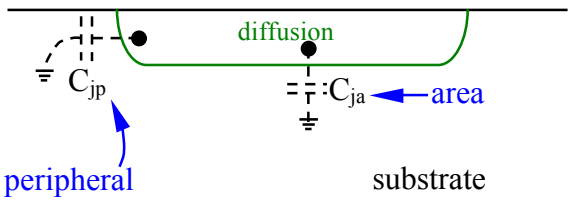
Hence,

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

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}	$0.1\,fF/\mu m^2$	$0.1\,fF/\mu m^2$
C_{jp0}	$0.9\,fF/\mu m$	$0.8\,fF/\mu m$

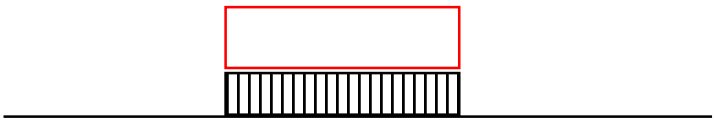
$$C_j = area * C_{ja0}\left(1-\frac{V_j}{\phi}\right)^{-ma} + perimeter * C_{jp0}\left(1-\frac{V_j}{\phi}\right)^{-mp}$$

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

ϕ = built-in junction potential $\approx$ 0.6-0.9 V

ma, mp = 0.3 ~ 0.5

Routing capacitance

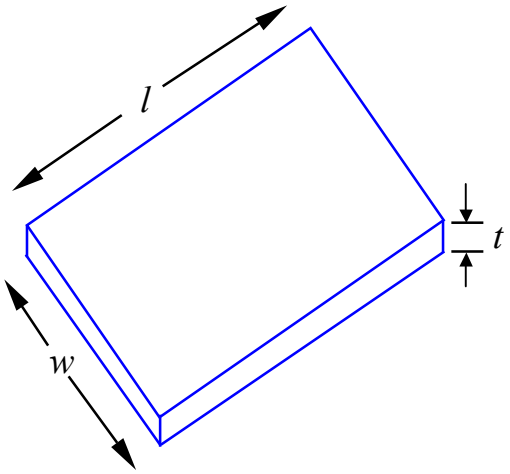


Representative CMOS Values

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

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

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

ρ_s: sheet resistivity

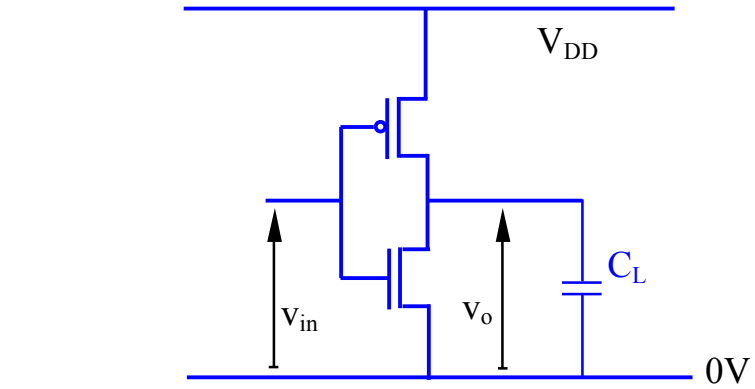
	sheet resistance $\Omega/sq.$		
material	min.	typical	max.
n well	1150	1300	1450
n^+	47	57	67
p^+	85	100	115
poly 1	17.5	22.5	27.5
poly 2	15	20	25
metal 1	0.066	0.072	0.078
metal 2	0.033	0.036	0.039

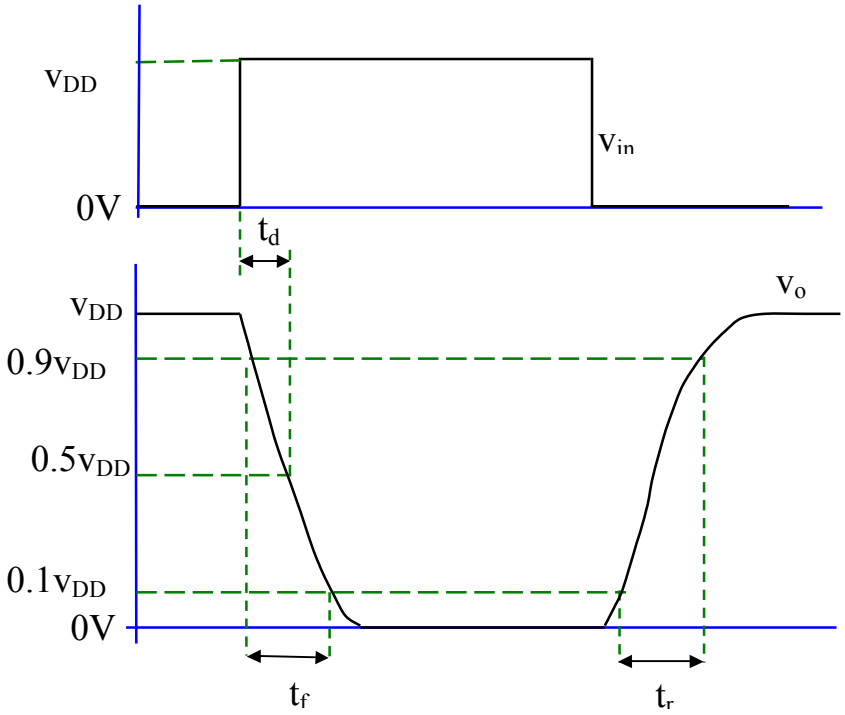
Sheet resistivity of transistor channel in the linear region

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

Timing Calculation

We will simplify this assuming a fixed load capacitance.



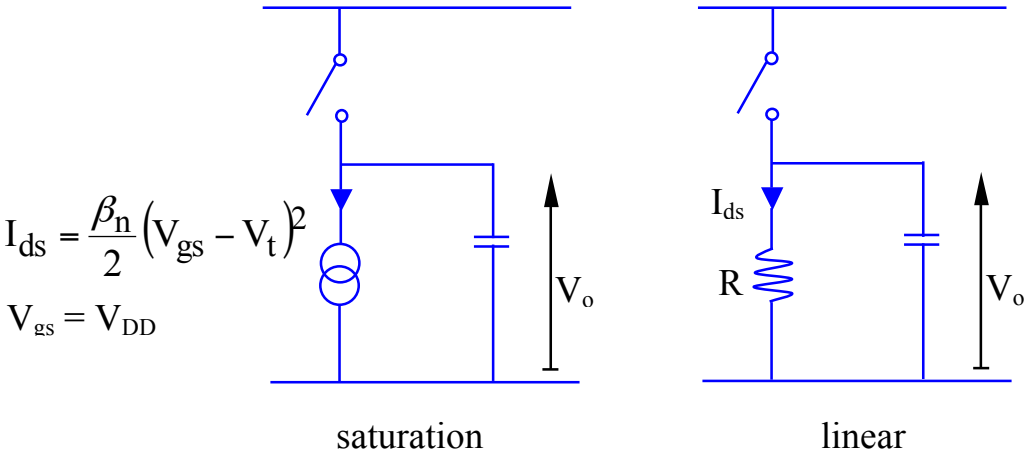


t_d : delay time

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



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

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

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

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

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 $0.1V_{DD}$

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

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

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

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

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

For $V_t = 0.2\, V_{DD}$

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

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

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 $V_{DD} - V_t$

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

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

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

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

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

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

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

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

How to reduce these delays and get faster circuits?

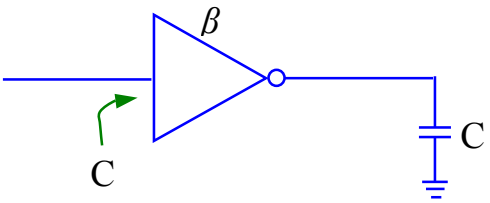
Hence for a fixed load, W of the devices can be increased to achieve the right required delay. Such fixed capacitances are normally dominated by the interconnect capacitance.

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.

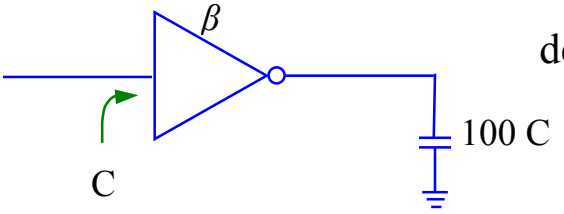
Driving intrinsic large capacitance load

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.



delay $= \tau_o =$ K C/β, where K is a constant.

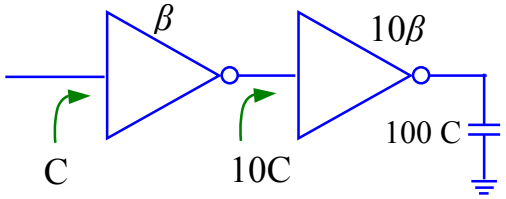


delay = K (100C)/β= 100 τ_o

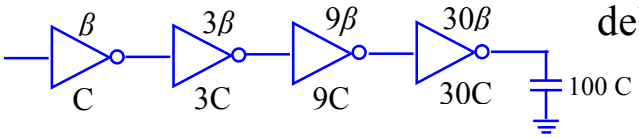
One may think that by increasing $\beta_n = \beta_p$ by a factor of 100 to 100 β will reduce the delay to τ_o. However, the capacitance seen at the input now becomes 100C as the gate area also increases by the factor 100 and delay at the input stage will now rise. Hence this basically transfers delay 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.

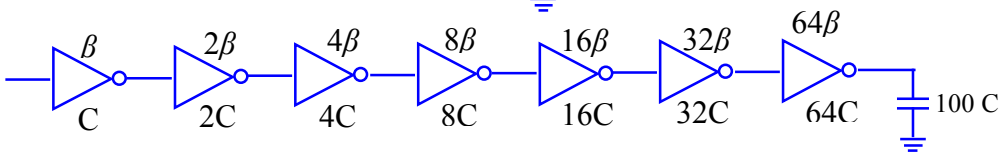
Here, for the last stage, the load is 100C, $\beta_n = \beta_p = 10\beta$. Hence the delay of the last stage = K*100C/(10β) = 10KC/β = 10 τ_o. For the first stage, load is 10C and $\beta_n = \beta_p = \beta$. Hence the delay of the first stage = K*(10C)/β = 10KC/β = 10 τ_o. The delay of every stage is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



delay = 10 τ_o + 10 τ_o
= 20 τ_o

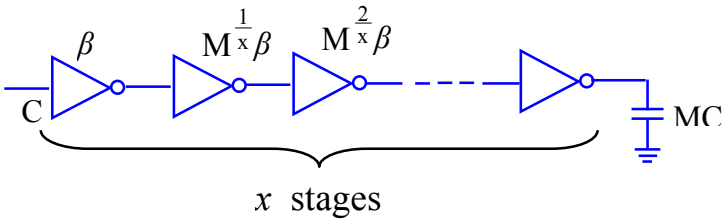


delay = (3 + 3 +3.3 +3.3)τ_o
= 12.6 τ_o



delay = (2 + 2 + 2 + 2 + 2 + 2 + 1.6) τ_o = 13.6 τ_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; $N_{a\ (or\ b)}$	α
	Supply voltage; V_{DD}	$1/\alpha$
	Electric field across gate oxide; E	1
	Depletion layer thickness; d	$1/\alpha$
	Parasitic capacitance; WL/t_{ox}	$1/\alpha$
	Gate delay; (VC/I)	$1/\alpha$
RESULTANT INFLUENCE	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.