UNIT - 01

CMOS Logic and Layouts

* Introduction and History:

- History:
  - 1958: First integrated circuit
    - by Jack Kilby at Texas Instruments.
  - 2003: Intel Pentium 4 microprocessor
    - contained 55 million transistors and 512-Mbit DRAM.

53% compound annual growth rate over 45 years, no other technology has grown so fast so long. This is due to steady miniaturization of transistors and improvement in manufacturing processes. Transistors have become smaller, faster, dissipate less power and cheaper to manufacture.

- Invention of Transistors:

During the first half of 20th century vacuum tubes were used which were large, expensive, power-hungry and unreliable.

- In 1947: First point contact transistor
  - by John Bardeen and Walter Brattain at Bell Labs
- Next Bell Labs developed bipolar junction transistors which were more reliable, less noisy and more power efficient. Small current into very thin base layer controls large currents between emitter and collector.
- By 1960's: Metal Oxide Semiconductor Field Effect Transistors (MOSFETs) They draw almost zero control current while idle. They are nMOS and pMOS using n-type and p-type silicon respectively.
- In 1963: First logic gates using MOSFET's
  - by Frank Wanlass at Fairchild. Fairchild's gates used both nMOS and pMOS transistors, earning the name complementary Metal Oxide Semiconductors (CMOS)

- In 1970's: nMOS transistors in processes became common. The were less expensive but still consumed power when idle.
- After 1980's: CMOS where considered for processes due to low consumption of power when idle.

+ Moore's Law:

In 1965, Gordon Moore observed that plotting the number of transistors that can be most economically manufactured on a chip gives a straight line on a semilogarithmic scale. He found transistor count doubled every 18 months. This observation has been called Moore's Law. The number of transistors in Intel Microprocessors has doubled every 26 months since the invention of 4004. Moore's law is driven primarily by scaling down the size of transistor and to a minor extent, by building larger chips.

The level of integration of chips has been classified into:

- small scale Integration (SSI): 10 gates
- Medium scale Integration (MSI): 1000 gates
- Large scale Integration (LSI): 10,000 gates
- Very large scale Integration (VLSI): > 10k gates.

A corollary of Moore's Law is Dennard's Scaling Law: as transistors shrink, they become faster, consume less power and are cheaper to manufacture.

★ VLSI Design Flow:

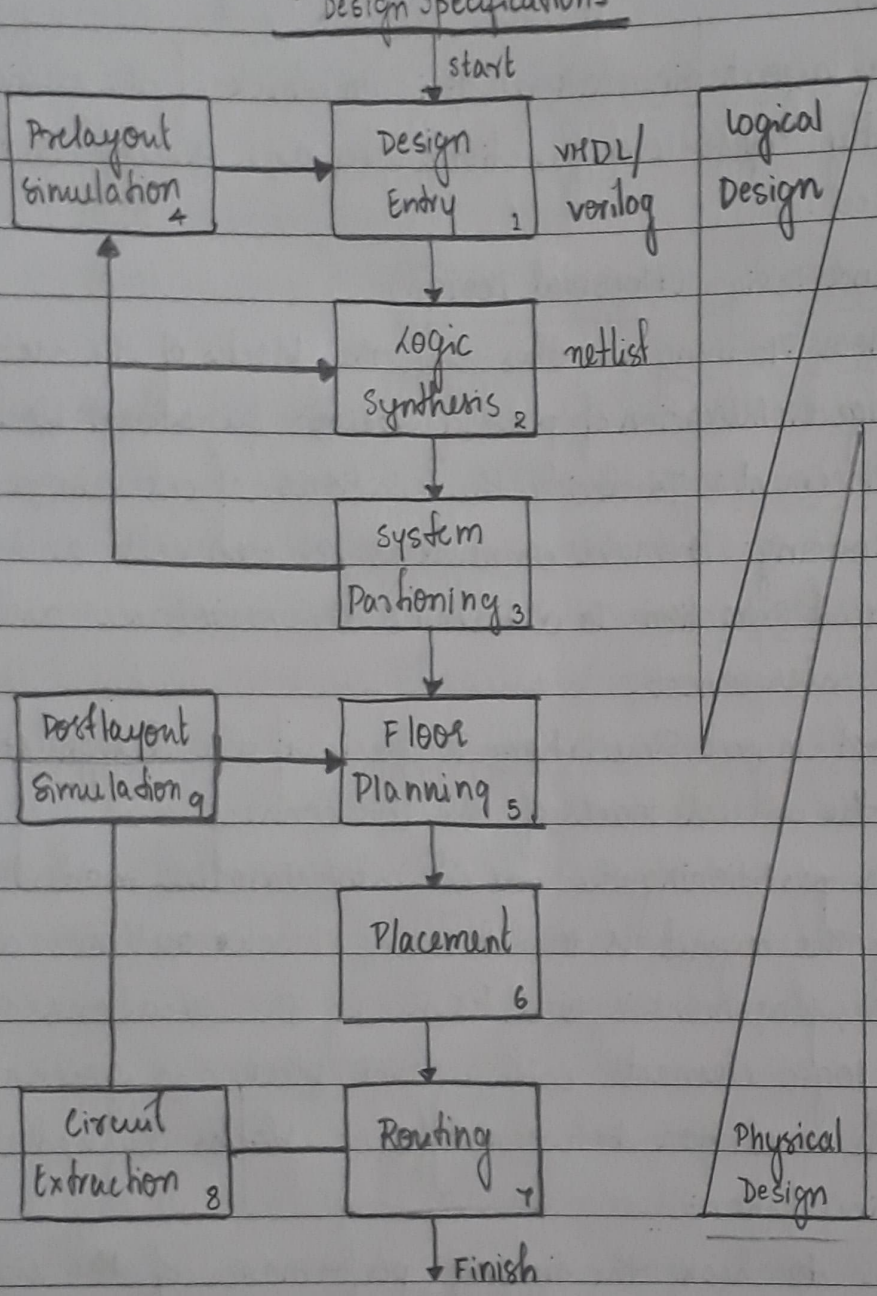
The steps to follow to design a VLSI chip are:

- Design specification:

The specifications to be mentioned in a VLSI Design are as follows:

1. The algorithm to be implemented in detail with mathematical representation: Determines complexity of design and gives an idea about the number of gates required.
2. Number of inputs and outputs in the design and number of bits in each of them: Determines number of pins to be used

3. Number of bits used in the internal arithmetic operation
4. Number of clock signals to be used in the design.
5. Maximum clock frequency to be used: Defines the speed of operation of the chip.
6. Area of the chip:
7. Power dissipation in the chip.



- **Front-end Design (Logic Design):**

1. **Design Entry:** The design is entered into an ASIC design system using a Hardware Description Language (HDL) or schematic Entry.
2. **Logic Synthesis:** Using HDL code netlist is generated i.e., logic cells and their connections.

Logic synthesis consists of following steps:

i. Technology Independent Logic optimization

ii. Translation: converting Behavioral description to structural domain.

iii. Technology mapping or library binding.

3. System Partitioning: It divides a large system into ASIC sized pieces.

4. Pre-layout simulation: It is to check if the design functions correctly. Gate-level functionality and timing details can be verified.

- Back-End Design (Physical Design):

5. Floor Planning: In this step the blocks of the netlist are arranged. Allocation of pins of various functional blocks

6. Placement: To decide the locations of cells in a block.

7. Routing: To make connections between cells and blocks.

8. Circuit Extraction: To determine the resistance and capacitance of the interconnect.

9. Post-layout simulation: To check if the design still works with the added loads of the interconnect.

After partitioning the circuit into smaller modules, floor planning the layout is to determine blocks outlines and pin locations, placement is to determining the locations of standard cells or logic elements within each block and routing is to establish connection between different blocks defined.

- Time Simulation:

It is to check the timing performance of the circuit such as setup and hold times of flip flops. It gives the detailed information of all gate delays and net delays of all the paths in the circuit to check whether they meet the timing constraints.

It gives an idea of maximum clock frequency used for the given routed design.

If the timing simulation fails, then first few iterations is rerouted to achieve a design with lesser delay. If rerouting fails, then design is re-entered by changing some serial processors to parallel processors to have faster operations.

Hence the trade of between hardware requirement (minimum) and clock speed (maximum).

- Fusing / Fabrication into the chip:

This is the last step in VLSI design. There are two types of design styles:

- Full custom Design : 1. ASIC
- Semi custom Design : 1. Cell based design
  2. Array based design (FPGA)

ASIC : Application Specific Integrated circuit are designed specifically for a given application. Full custom designs involve the complete design to be hand-crafted in transistor level so as to optimize the circuit for performance and are for a given application.

Cell based designs use libraries of predesigned cells which are then placed and wired to complete the design.

Array Based designs use a prefabricated matrix of non-connected components. FPGA : Field Programmable Gate Array uses programmable logic modules and also programmable interconnections in which configuration data is loaded during each application.

- Fabrication, Packaging and Testing:

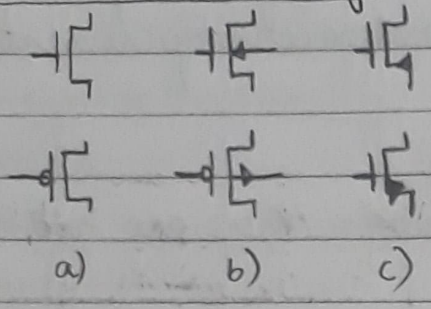
Once a chip design is complete, it is taped out for manufacturing. Tapeout gets its name from the old practice of writing a specification of masks to magnetic tape; now the mask descriptions are usually sent to the manufacturer electronically.

Masks are made by etching a pattern of chrome on glass with an electron beam. A set of masks for a modern process can be expensive.

★ MOS Transistor Theory:

MOS Transistor was introduced in terms of its operation as an ideal switch.

- MOS Transistor symbols

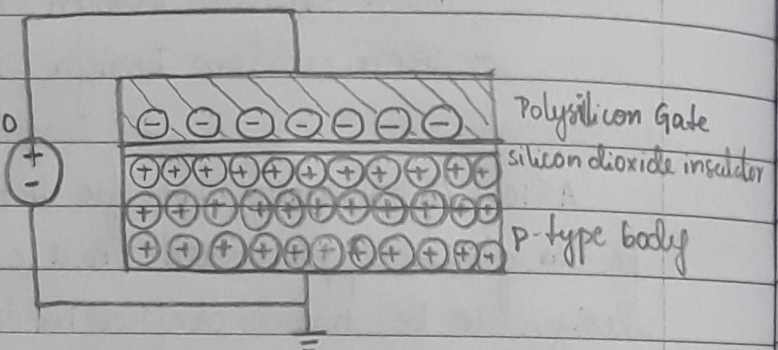


The MOS transistor is a majority-carrier device in which current is a conducting channel between the source and drain is controlled by a voltage applied to the gate.

- MOS Structure

Figure a)

A negative voltage is applied to the gate, so there is a negative charge on the gate.



The mobile positively charged holes are attracted to the region beneath the gate. This is called the accumulation mode.

Figure b)

A low positive voltage is applied to the gate, resulting in some positive charges on the gate. The holes in the body is repelled from the region directly beneath the gate, resulting in a depletion region forming below the gate.

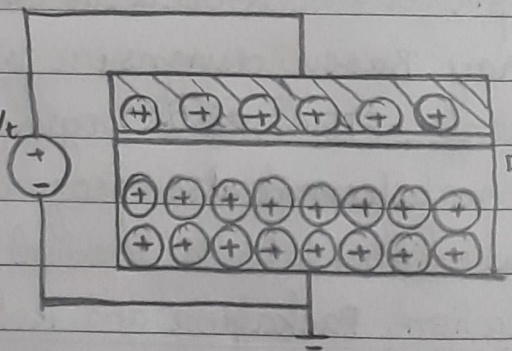
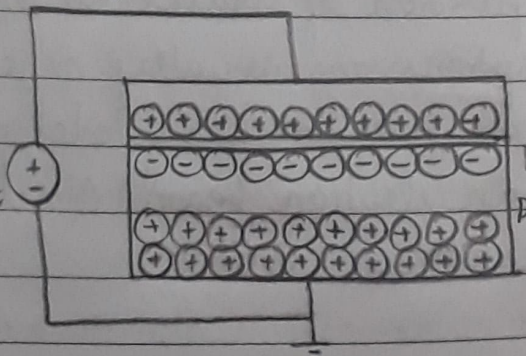


Figure c)

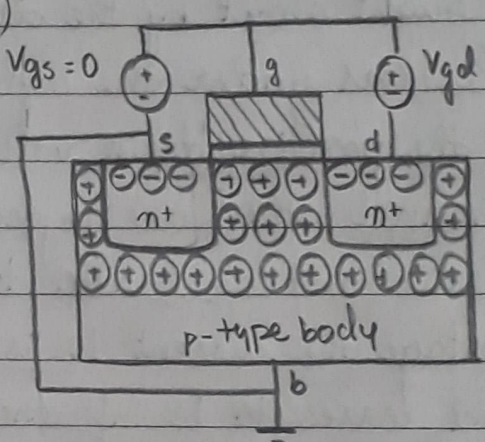
A higher positive potential exceeding the critical threshold voltage V_t is applied, attracting more positive charge to the gate. The holes are repelled further and a small number of



free electrons in the body are attracted to the region beneath the gate. This conductive layer of electrons in the p-type body is called the inversion layer. The threshold voltage depends on the number of dopants in the body and thickness of the oxide (t_{ox}). It is usually positive but can be engineered to be negative.

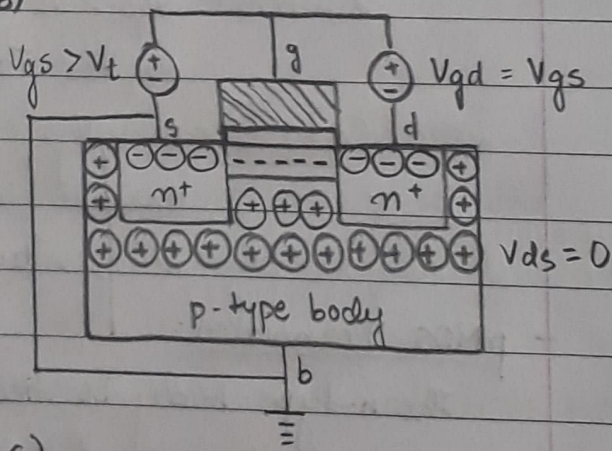
- nMOS Transistor:

a) The gate to source voltage V_{gs} is less than the threshold voltage. The source and drain have free electrons. The body has free holes but no free electrons. The junction between the body and source or drain are reverse-biased, so almost zero current flows. This mode of operation is called cutoff.

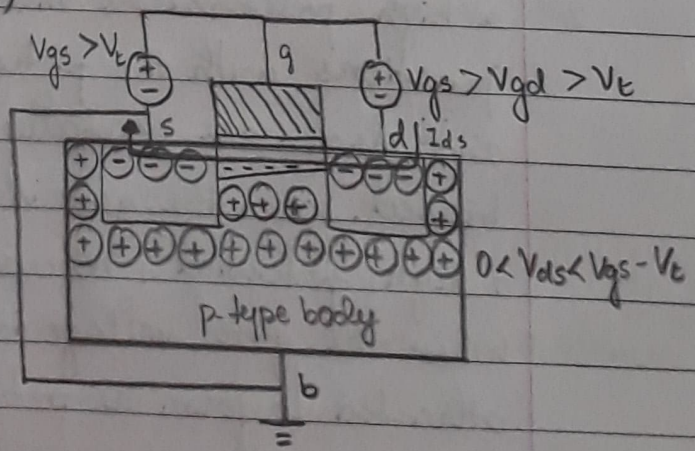


b) The gate voltage is greater than the threshold voltage. An inversion region of electrons (majority carriers) called the channel connects the source and drain creating a conducting path. The number of carriers and conductivity increases with the gate voltage.

$V_{ds} = V_{gs} - V_{gd}$ If $V_{ds} = 0$, there is no electric field tending to push current from drain to source.

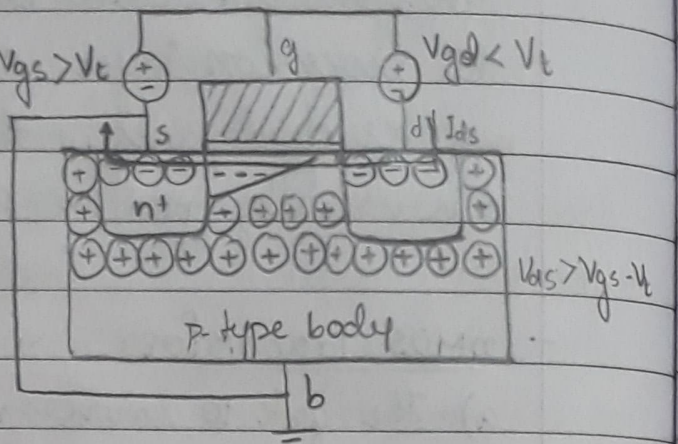


c) When a small positive potential V_{ds} is applied to the drain, current I_{ds} flows through the channel from drain to source. In linear mode of operation, current



current increases with both the drain and gate voltage.

d) If V_{ds} becomes sufficiently large that $V_{gd} < V_t$, the channel is no longer inverted near the drain and becomes pinched off. However the conduction is still brought about by the drain of electrons under the influence of the positive drain voltage. As electrons reach the end of the channel, that are injected into the depletion region near the drain and accelerated towards the drain. Above this drain voltage the current I_{ds} is controlled only by the gate voltage and ceases to be influenced by the drain. This mode is called *saturation*.



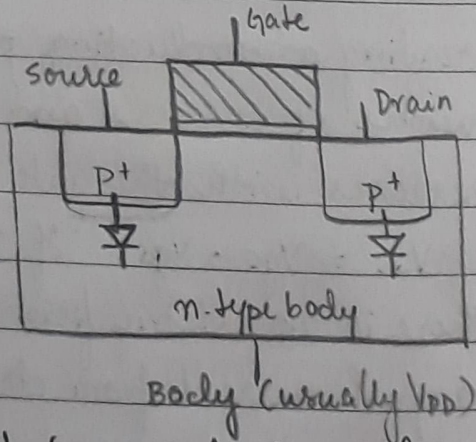
∴ Cut off : no current flows : $V_{gs} < V_t$

Linear : $V_{gs} > V_t$ and V_{ds} is small. Current flow is proportional to V_{ds}. Transistor acts as linear resistor

Saturation : $V_{gs} > V_t$ and V_{ds} is large. Transistor acts as a current source and current flow becomes independent of V_{ds}.

- pMOS Transistor:

The n-type body is tied to a high potential so that the junctions with the p-type source and drain are normally reverse biased. When the gate is also at a high potential, no current flows between drain and source. When the gate voltage is lowered by a threshold V_t, holes are attracted to form a p-type channel beneath the gate, allowing current to flow between drain and source. The threshold voltages of the two types of transistors are not necessarily equal, so we use the terms V_{tn} and V_{tp} to distinguish the nMOS and pMOS thresholds

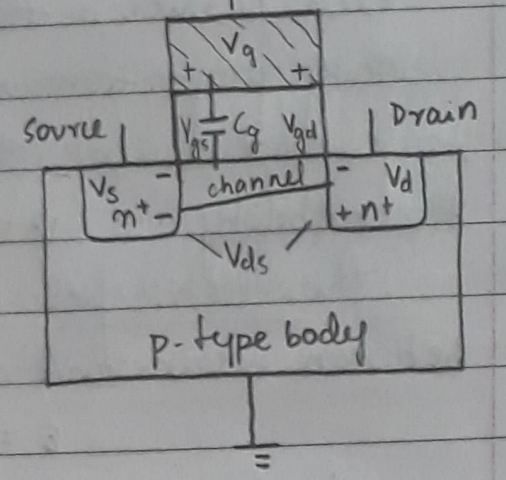


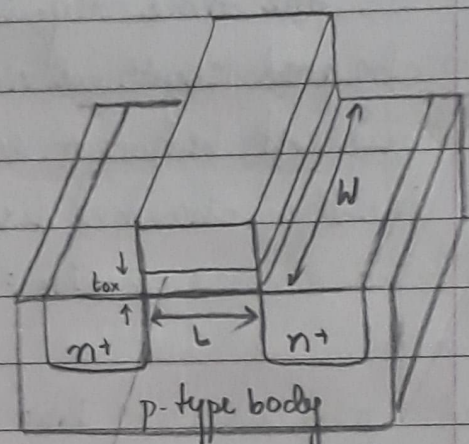
Keerthana Ashok

* Ideal IV characteristics:

In the cut off region, there is no channel and almost zero current flows from drain to source.

In the other regions, the gate attracts carriers (electrons) to form a channel. The electron drift from source to drain at a rate proportional to the electrical field between these regions. Thus we can compute currents if we know the amount of charge in the channel and the rate at which it moves.





SiO_2 Gate Oxide
(Good insulator, $\varepsilon_{ox} = 3.9\varepsilon_0$)

wkt, charge on each plate of capacitor is $Q = CV$.

Therefore,

$$Q_{channel} = C_g (V_{gc} - V_t)$$

where C_g: capacitance of the gate to the channel

$V_{gc} - V_t$: amount of voltage attracting charge to the channel beyond the minimum required to invert from p to n.

The gate voltage is referenced to the channel which is not grounded. If the source is at V_s and the drain is at V_d, the average is $V_c = \frac{V_s + V_d}{2} = V_s + \frac{V_{ds}}{2}$

$$\therefore V_{gc} = V_{gs} - \frac{V_{ds}}{2}$$

If the gate has length L and width W and the oxide thickness is t_{ox}, the capacitance is:

$$C_g = \varepsilon_{ox} \frac{WL}{t_{ox}} \quad \text{where} \quad \frac{\varepsilon_{ox}}{t_{ox}} = C_{ox}$$

Each carrier in the channel is accelerated to an average velocity proportional to the lateral electric field, ie, between source and drain. The constant of proportionality μ is called the mobility: $v = \mu E$

The electric field E is the voltage difference between between drain and source.

$$E = V_{ds}/L$$

The time required for carriers to cross the channel is the channel length divided by the carrier velocity. Therefore the current between source and drain is the total amount of charge in the channel divided by the time required to cross

$$I_{ds} = \frac{Q_{channel}}{L/v}$$

$$I_{ds} = \mu C_{ox} \frac{W}{L} \left(V_{gs} - V_t - \frac{V_{ds}}{2} \right) V_{ds}$$

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

where $\beta = \mu C_{ox} \frac{W}{L}$

If $V_{ds} > V_{dsat} = V_{gs} - V_t$, the channel is no longer inverted in the vicinity of the drain, we say it is pinched off. Beyond this point, called the drain saturation voltage, increasing the drain voltage has no further effect on current

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

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

Therefore;

$$I_{ds} = \begin{cases} 0 & V_{gs} < V_t \quad \text{cut off} \\ \beta \left(V_{gs} - V_t - \frac{V_{ds}}{2} \right) V_{ds} & V_{ds} < V_{dsat} \quad \text{linear} \\ \frac{\beta}{2} (V_{gs} - V_t)^2 & V_{ds} > V_{dsat} \quad \text{saturation} \end{cases}$$

★ Simple MOS capacitance Models:

The gate of an MOS transistor is a good capacitor. Indeed, its capacitance is necessary to attract charge to invert the channel, so high gate capacitance is required to obtain high I_{ds}. The gate capacitor can be viewed as a parallel plate capacitor with the gate on top and channel on bottom with the oxide dielectric between.

$$\therefore C_g = C_{ox} WL$$

★ Detailed MOS Gate capacitance Model:

Gate to channel capacitance creates channel charge neccessary for operation. (intrinsic)

Source and drain have capacitance to body (parasitic capacitance) across reverse biased diodes called diffusion capacitance because it is associated with source/drain diffusion.

Let $C_{ox} WL = C_0$

When the transistor is on, the channel extends from the source to drain (if the transistor is unsaturated or to the pinchoff point otherwise)

$$C_g = C_{gb} + C_{gs} + C_{gd}$$

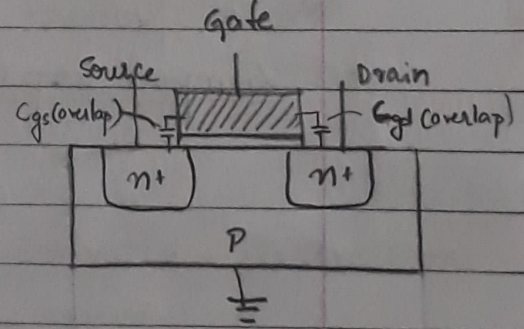
C_{gb} : gate to body
C_{gs} : gate to source
C_{gd} : gate to drain

Parameter	Cut off	Linear	saturation
C_{gb}	C_0	0	0
C_{gs}	0	$C_0/2$	$2/3\ C_0$
C_{gd}	0	$C_0/2$	0
$C_g = C_{gs} + C_{gd} + C_{gb}$	C_0	C_0	$2/3\ C_0$

Approximation of intrinsic MOS gate capacitance

In reality the gate overlaps source and drain. Thus, the gate capacitance should include not only the intrinsic capacitance but also parasitic overlap capacitances.

$$C_{gs}(\text{overlap}) = C_{ox} WL_D$$
$$C_{gd}(\text{overlap}) = C_{ox} WL_D$$



overlap capacitances

Parameter	Cutoff	Linear	Saturation
C_{gb} (total)	C_o	0	0
C_{gd} (total)	$C_{ox} WL_D$	$C_o/2 + C_{ox} WL_D$	$C_{ox} WL_D$
C_{gs} (total)	$C_{ox} WL_D$	$C_o/2 + C_{ox} WL_D$	$2/3 C_o + C_{ox} WL_D$

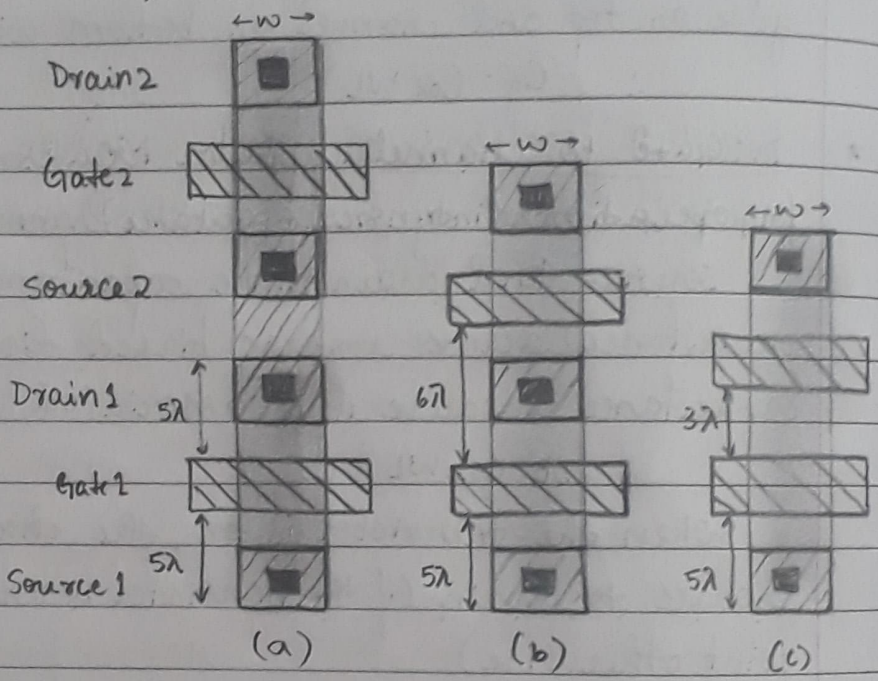
Detailed Gate Capacitance

* **Detailed MOS Diffusion Capacitance Model:**

a) Each source and drain has its own isolated region of contacted diffusion.

b) the drain of the bottom transistor and source of the top transistor form a shared contacted diffusion region.

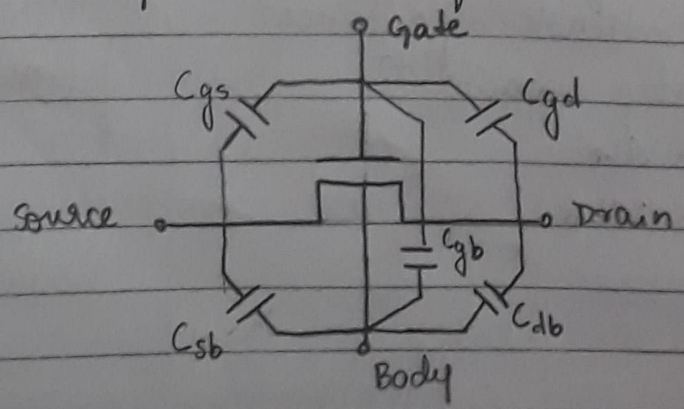
c) the source and drain are merged into an uncontacted region.



Diffusion region geometries

Here C_{sb} and C_{db} are undesired capacitance (parasitic) due to the reverse biased p-n junctions between source diffusion and body and drain diffusion and body. capacitance depends on area and perimeter, so use small diffusion nodes comparable to C_g for contacted diffusion and $1/2\ C_g$ for uncontacted. It varies with process.

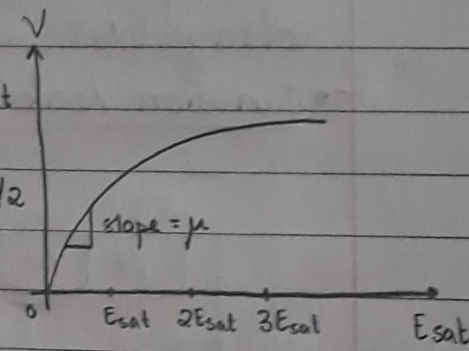
• **Lumped representation of MOSFET capacitances**



★ Non-ideal I-V effects:

- The saturation current increases less than quadratically with increasing V_{gs}. This is caused by two effects:
  - Velocity saturation:

At strong lateral fields resulting from high V_{ds}, carrier velocity ceases to increase linearly with electric field and results in lower I_{ds} than expected.

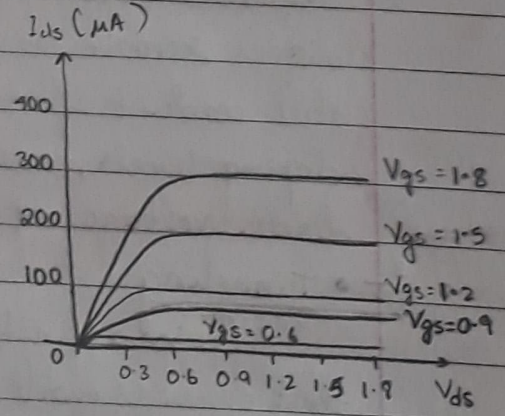


Carrier Velocity vs electric field

  - Mobility Degradation: Strong vertical fields resulting from large V_{gs} cause the carriers to scatter against the surface and also reduce the carrier mobility. This effect is called mobility degradation.

- Channel length Modulation:

The reverse biased p-n junction between the drain and the body forms a depletion region with length L' that increases with V_{db}. The depletion region effectively shorten the channel length to: $L_{eff} \cong L - L'$



I-V characteristics of nMOS transistor with channel length modulation.

Assuming the source voltage is close to the body voltage $V_{db} \sim V_{sb}$. Hence, increasing V_{ds} decrease the effective channel length. Shorter channel length results in higher current.

- Body Effect:

The potential difference between source and body V_{sb} affects (increases) the threshold voltage. The threshold voltage depends on V_{sb}, process, doping, temperature.

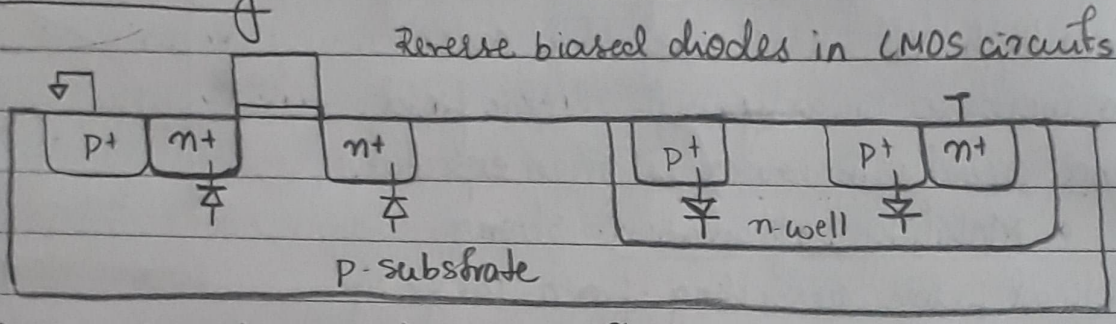
leakage currents

- Subthreshold conduction:

The ideal transistor I-V model assumes current only flows from drain to source when $V_{gs} > V_t$. In real transistors (OFF mode),

current doesn't abruptly cut off below threshold but rather drop off exponentially. This leaking current when the transistor is nominally OFF depends on: process (ε_{ox}, t_{ox}), doping levels, device geometry (W, L), temperature (T) and subthreshold voltage (V_t).

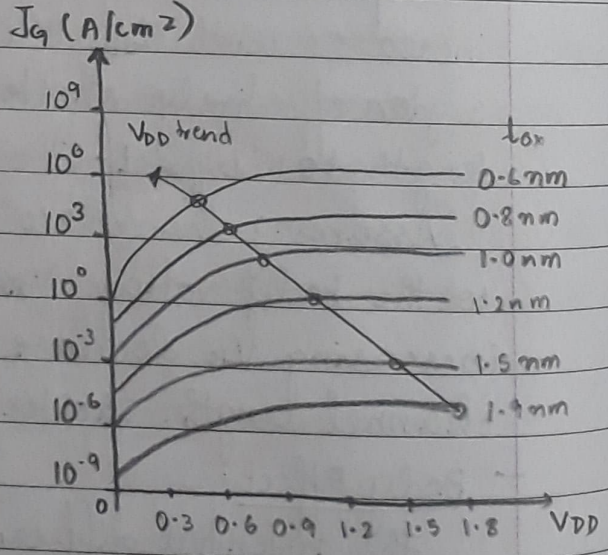
- Junction leakage:



The p-n junctions between diffusion and the substrate or well form diodes. The well-to-substrate in another diode. Substrate and well are tied to GND and VDD to ensure these diodes remain reverse biased. But, reverse biased diodes still conduct a small amount of current that depends on: doping levels, area and perimeter of the diffusion region, the diode voltage.

- Tunneling:

There is a finite probability that carriers will tunnel through the gate oxide. This results in gate leakage current flowing into the gate. The probability drops off exponentially with t_{ox}. For oxides thinner than 15 - 20 Å, tunneling becomes a factor.



Gate leakage current

- Temperature Dependence:

Transistor characteristics are influenced by temperature

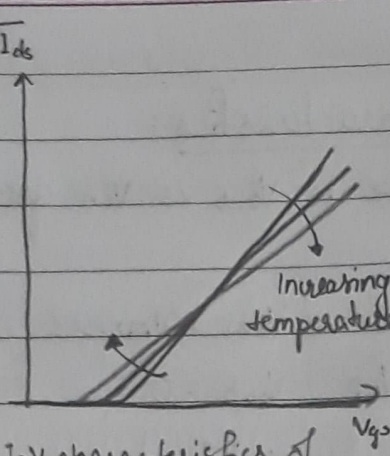
- μ decreases with T
- V_t decreases linearly with T
- $I_{leakage}$ increases with T.

Date:

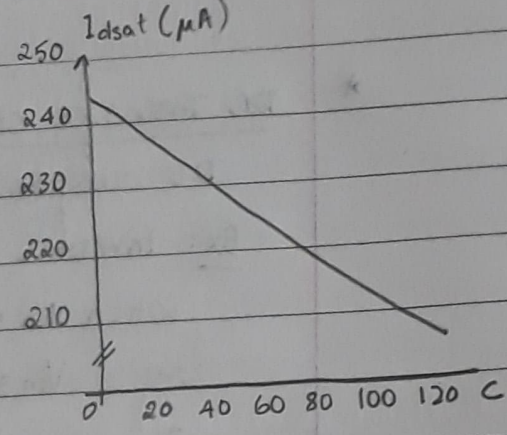
Keerthana Ashok

- ON current decreases with T
- OFF current increases with T

Thus circuit performances are worst at high temperature.



I-V characteristics of nMOS transistor in saturation at various temperatures



I_{dsat} vs temperature

- Geometry Dependence:

Layout designers draw transistors with W_{drawn}, L_{drawn}. Actual dimensions may differ from some factor X_W and X_L. The source and drain tend to diffuse laterally under the gate by L_D, producing a shorter effective channel. Similarly, diffusion of the bulk by W_D decreases the effective channel width. In process below 0.25μm the effective length of the transistor also depends significantly on the orientation of the transistor.

- Impacts of non-ideal W effects:
  - Threshold is a significant fraction of the supply voltage
  - Leakage is increased causing gates to consume power when idle and limits the amount of time that data is retained. Leakage increases with temperature.
  - velocity saturation and mobility degradation result in less current than expected at high voltage thus no point in trying to use high V_{DD} to achieve fast transistors.
  - Transistors in series partition the voltage across each transistor thus experience less velocity saturation. Thus they tend to be faster than a single transistor.
  - If two transistors should behave identically, both should have the same dimensions and orientations.

* DC Transfer characteristics:

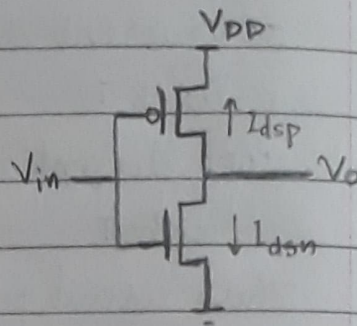
DC response is Vout vs Vin for a gate.

Ex: Inverter

When Vin = 0 Vout = VDD

Vin = VDD Vout = 0

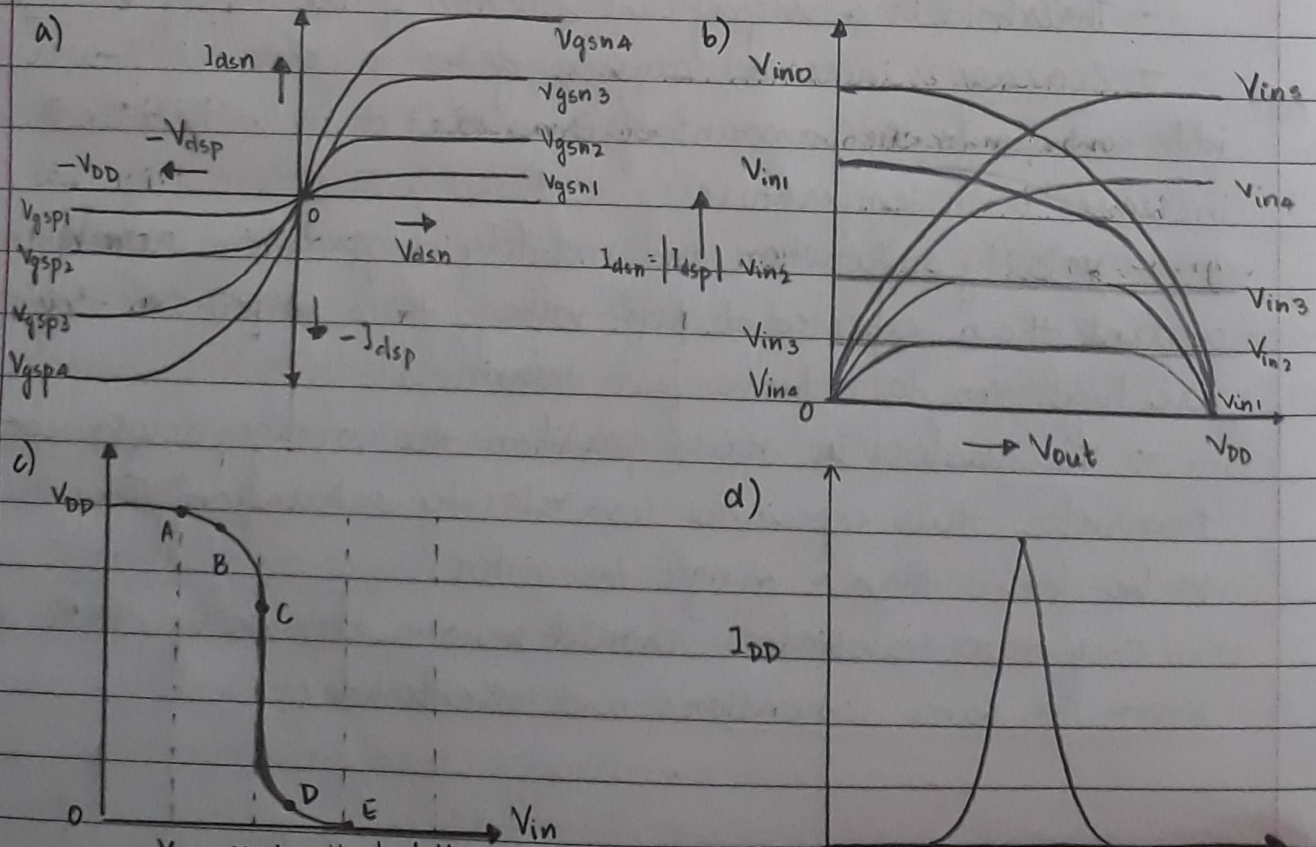
By KCL, $I_{dsn} = |I_{dsp}|$



A CMOS inverter.

	Cutoff	Linear	saturated	
nMOS	$V_{gsn} < V_{tn}$ ($V_{in} < V_{tn}$)	$V_{gsn} > V_{tn}$ ($V_{in} > V_{tn}$) $V_{dsn} > V_{gsn} - V_{tn}$ ($V_{out} > V_{in} - V_{tn}$)	$V_{gsn} > V_{tn}$ ($V_{in} > V_{tn}$) $V_{dsn} > V_{gsn} - V_{tn}$ ($V_{out} > V_{in} - V_{tn}$)	$V_{gsn} = V_{in}$ $V_{dsn} = V_{out}$
pMOS	$V_{gsp} > V_{tp}$ ($V_{in} > V_{tp} + V_{DD}$)	$V_{gsp} < V_{tp}$ ($V_{in} < V_{tp} + V_{DD}$) $V_{dsp} > V_{gsp} - V_{tp}$ ($V_{out} > V_{in} - V_{tp}$)	$V_{gsp} < V_{tp}$ ($V_{in} < V_{tp} + V_{DD}$) $V_{dsp} < V_{gsp} - V_{tp}$ ($V_{out} < V_{in} - V_{tp}$)	$V_{gsp} = V_{in} - V_{DD}$ $V_{dsp} = V_{out} - V_{DD}$ $V_{tp} < 0$

Relationships between voltages for the three regions of operation of a CMOS inverter



Keerthana Ashok

Graphical derivation of CMOS inverter DC characteristics:

a) IV characteristics

PMOS is wider than NMOS such that $\beta_n = \beta_p$

b) Load line analysis

For given V_{in}: Plot I_{dsn}, I_{dsp} vs V_{out}

V_{out} must be where |currents| are equal

c) operating regions

A: nmos: cutoff
pmos: linear

B: nmos: saturation
pmos: linear

C: nmos: saturation
pmos: saturation

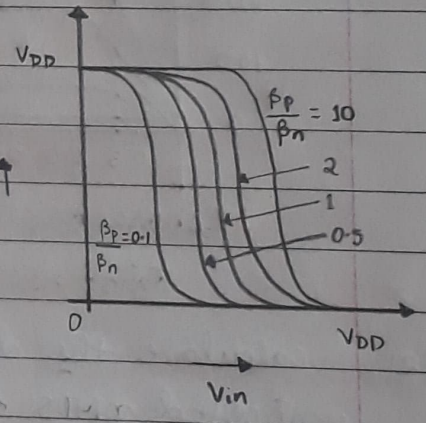
D: nmos: linear
pmos: saturation

E: nmos: linear
pmos: cutoff

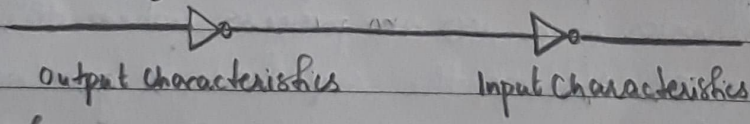
Beta Ratio:

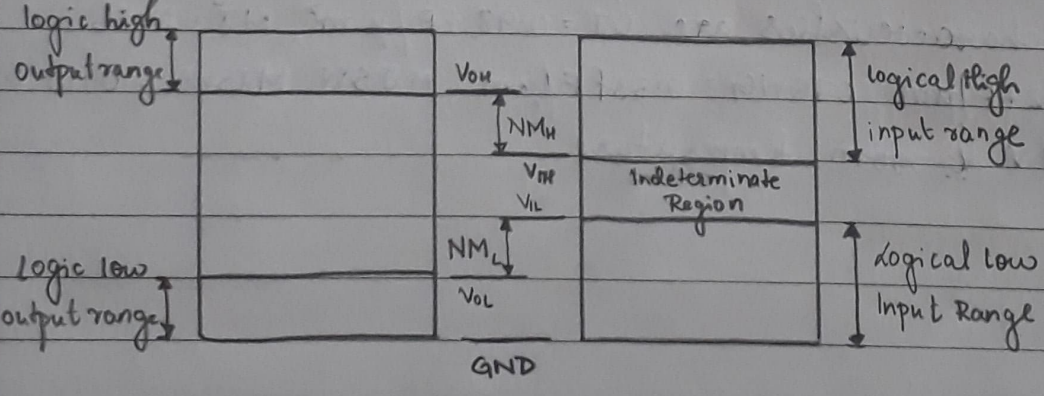
If $\frac{\beta_p}{\beta_n} \neq 1$, switching point will more from $V_{DD}/2$ and are called skewed gate.

other gates collapse into equivalent matter.



Noise Margins:





NM_L: low noise margin
NM_H: high noise margin
V_{IH} = minimum HIGH input voltage
V_{IL} = maximum LOW input voltage
V_{OH} = minimum HIGH output voltage
V_{OL} = maximum LOW output voltage

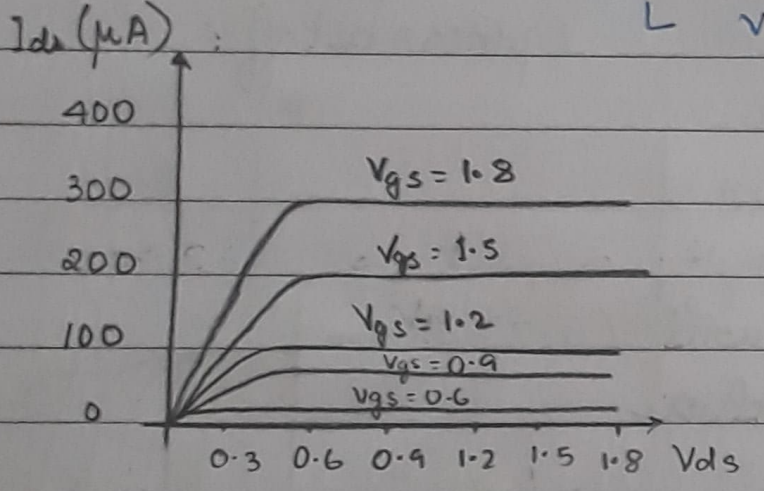
This parameter allows to determine the allowable noise voltage on the input of the gate so that the output will not be corrupted

Q1. Consider an nMOS transistor in a 180nm process with W/L = 4/2λ (i.e., 0.36/0.18μm). In this process, the gate oxide thickness is 40 Å and the mobility of electrons is 180 cm²/V·s at 70°C. The threshold voltage is 0.4V. Plot Ids vs Vds for Vgs = 0, 0.3, 0.6, 0.9, 1.2, 1.5 and 1.8 V.

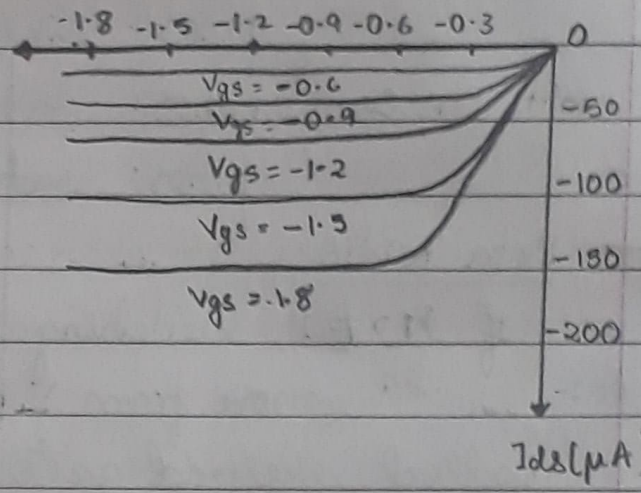
— We calculate β

$$\beta = \mu C_{ox} \frac{W}{L} = \left(180 \frac{cm^2}{V \cdot s}\right)\left(\frac{3.9 \times 8.85 \times 10^{-14}\ F/cm}{40 \times 10^{-8}\ cm}\right)\left(\frac{W}{L}\right)$$

$$= 155 \frac{W}{L} \frac{\mu A}{V^2}$$



I V characteristics of ideal nMOS transistor



I V characteristics of ideal pMOS transistor