METHODIST
COLLEGE OF ENGINEERING AND TECHNOLOGY
Approved by AICTE New Delhi | Affiliated to Osmania University, Hyderabad
Estd : 2008 Address : King Koti Road, Abids, Hyderabad, Telangana, 500001 | Email : principal@methodist.edu.in

DEPARTMENT OF

ELECTRONICS AND COMMUNICATION ENGINEERING

# LECTURE NOTES

# ON

# VLSI DESIGN

## B.E VII Semester (PC702 EC)

**Mr. I. SRIKANTH,**
**Associate Professor**
**Department of ECE**

**2019-2020**

# VLSI DESIGN

## INTRODUCTION TO IC TECHNOLOGY

Over the last two decades electronics industry has achieved remarkable growth, mainly due to the advent of **Very-large-scale integration** (**VLSI**).VLSI is the process of creating an integrated circuit( IC) by combining thousands of transistors into a single chip. The number of applications of IC's is in high performance computing, telecommunications, consumer electronics etc. The required computational power (or the intelligence) of these applications is the driving force the fast development of this field.

As more and more complex functions are required in various data processing and telecommunications devices, the need to integrate these functions in a small system/package is also increasing. The levels of integration are measured by the no. of logic gates in a monolithic chip. Table 1.1 shows evaluation of logic complexity in integrated circuits.

| Classification | No. of active devices per chip |
|---|---|
| Small Scale Integration(SSI) | 1-100 |
| Medium Scale Integration(MSI) | 100-1000 |
| Large Scale Integration(LSI) | 1000-10000 |
| Very Large Scale Integration(VLSI) | 10exp4-10exp5 |
| Ultra Large Scale Integration(ULSI) | 10exp5-10exp6 |
| Super Large Scale Integration(SLSI) | 10exp6-10exp7 |
| Extra Large Scale Integration(ELSI) | 10exp7-10exp8 |
| Giga Scale Integration(GSI) | >10exp8 |

Table:1.1 Evaluation of logic complexity in integrated circuits

A measure of progress of IC's is determined by the no.of devices per chip as well as the size of the chip and the process technology used within. The continued trends have been to produce smaller, faster, more reliable and less expensive systems which consume less power. Table 1.2 shows the evaluation of process technology in integrated circuits.

| Year | Technology |
|---|---|
| 1971 | 10 µm |
| 1974 | 6 µm |

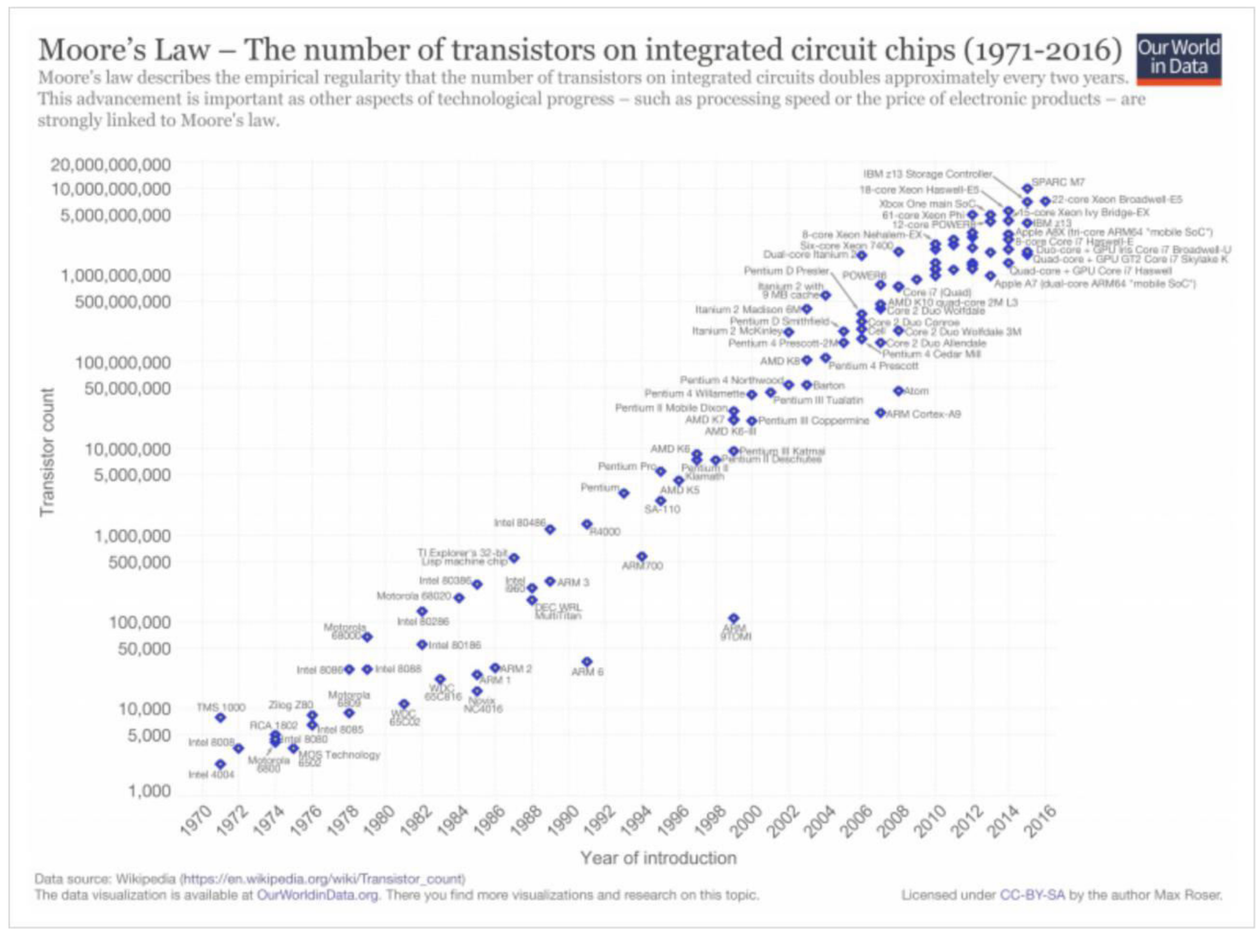
| | |
|---|---|
| 1977 | 3 µm |
| 1982 | 1.5 µm |
| 1985 | 1µm |
| 1989 | 800 nm |
| 1994 | 600 nm |
| 1995 | 350 nm |
| 1997 | 250 nm |
| 1999 | 180 nm |
| 2001 | 130 nm |
| 2004 | 90 nm |
| 2006 | 65 nm |
| 2008 | 45 nm |
| 2010 | 32 nm |
| 2012 | 22 nm |
| 2014 | 14 nm |
| 2017 | 10 nm |
| 2018 | 7 nm |
| 2020 | 5 nm |

Next technology nodes are $36A^0$,$25A^0$,$18A^0$, $13A^0$, $9A^0$ ($1A^0$ = nm)

**The Integrated Circuit (IC) era:**

Such has been the potential of the silicon integrated circuit that there has been an extremely rapid growth in the number of transistors (as a measure of complexity) being integrated into circuits on a single silicon chip. The relationship between the no. of transistors per chip versus the year has become known as 'Moor's first law' after declaration made by Gordon moor in the 1960s.

**Moore's law** is the observation that over the history of computing hardware, the number of transistors on integrated circuits doubles approximately every two years. The period often quoted as "18 months" is due to Intel executive David House, who predicted that period for a doubling in chip performance.



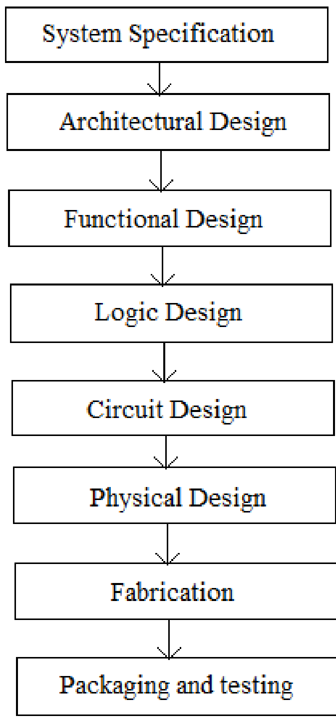

Although over the past several years, Silicon CMOS technology has become the dominant fabrication process for relatively high performance and cost effective VLSI circuits, the revolutionary nature of new systems such as the wired and wireless communication technologies, high performance imaging systems, smart appliances and the like are constantly challenging the boundaries of various technological fronts including silicon CMOS. The processing requirements for the image capture, conversion, compression, decompression, enhancement and display of increasingly higher quality multimedia content and future generation multimedia, together with the emergence of new and complex optical and photonics technologies being driven by microelectronics, place heavy demands on current standard CMOS technology integrated systems, particularly when low power and high performance solutions are required.

Although technology is continuously evolving to produce smaller systems with minimized power dissipation, the IC industry is facing major challenges due to constraints on

power density (W/cm2) and high dynamic (operating) and static (standby) power dissipation. The key to overcome these challenges lies in improvements in design, material and manufacturing processes. The significant issues that relate to successful designs include (a) approach to system design cycle (b) workable transistors models.

**System Design Cycle:**

The VLSI design cycle starts with a formal specification of a VLSI chip, follows a series of steps, and eventually produces a packaged chip.




VLSI design Flow

**1.System Specification:**

The first step of any design process is to set the specifications of the system. System specification is a high level representation of the system. The factors to be considered in this process are performance, functionality, and physical dimensions (size of the die (chip)). The fabrication technology and design techniques are also considered. The specification of a system is a compromise between market requirements, technology and economical viability.

**2.Architectural Design:**

The basic architecture of the system is designed in this step. The architectural design of a VLSI circuit begins with the development of the idea of the main module that will be followed by the definition of the module in terms of inputs, outputs, and a description of the specific function. This also includes number of ALUs, Floating Point units, number and structure of pipelines, and size of caches among others.

**3. Functional Design:**

In this step, main functional units of the system are identified. This also identifies the interconnect requirements between the units. The area, power, and other parameters of each unit are estimated and functional aspects of the system are considered here.

For example, it may specify that a multiplication is required, but exactly in which mode such multiplication may be executed is not specified. We may use a variety of multiplication hardware depending on the speed and word size requirements. The key idea is to specify behavior, in terms of input, output and timing of each unit, without specifying its internal structure.

The outcome of functional design is usually a timing diagram. This information leads to improvement of the overall design process and reduction of the complexity of subsequent phases.

**4. Logic Design:**

In this step the control flow, word widths, register allocation, arithmetic operations, and logic operations of the design that represent the functional design are derived and tested.

This description is called Register Transfer Level (RTL) description. RTL is expressed in a Hardware Description Language (HDL), such as VHDL or Verilog. This description can be used in simulation and verification. This description consists of Boolean expressions and timing information. The Boolean expressions are minimized to achieve the smallest logic design which conforms to the functional design. This logic design of the system is simulated and tested to verify its correctness. In some special cases, logic design can be automated using *high level synthesis* tools. These tools produce a RTL description from a behavioral description of the design.

**5. Circuit Design:**

The purpose of circuit design is to develop a circuit representation based on the logic design. The Boolean expressions are converted into a circuit representation by taking into consideration the speed and power requirements of the original design. *Circuit Simulation* is used to verify the correctness and timing of each component.

The circuit design is usually expressed in a detailed circuit diagram. This diagram shows the circuit elements (cells, macros, gates, transistors) and interconnection between these elements. This representation is also called a *netlist.* Tools used to manually enter such description are called *schematic capture tools.* In many cases, a netlist can be created automatically from logic (RTL) description by using *logic synthesis* tools.

**Physical Design**:

In this step the netlist converted into a geometric representation. This geometric representation of a circuit is called a *layout.* Layout is created by converting each logic component (cells, macros, gates, transistors) into a geometric representation which performs the intended logic function of the corresponding component. Connections between different components are also expressed as geometric patterns typically lines in multiple layers.

The exact details of the layout also depend on design rules, which are guidelines based on the limitations of the fabrication process and the electrical properties of the fabrication materials. Physical design is a very complex process and therefore it is usually broken down into various sub-steps. In many cases, physical design can be completely or partially automated and layout can be generated directly from netlist by *Layout Synthesis* tools. Various verification and validation checks are performed on the layout during physical design.

**7. Fabrication:**

After layout and verification, the design is ready for fabrication. Since layout data is typically sent to fabrication on a tape, the event of release of data is called *Tape Out.* Layout data is converted into photo-lithographic masks, one for each layer. Masks identify spaces on the wafer, where certain materials need to be deposited, diffused or even removed. Silicon crystals are grown and sliced to produce wafers. The fabrication process consists of several steps involving deposition, and diffusion of various materials on the wafer. During each step one mask is used. Several dozen masks may be used to complete the fabrication process.

**8. Packaging, Testing and Debugging**:

Finally, the wafer is fabricated and cut into individual chips in a fabrication process. Each chip is then packaged and tested to ensure that it meets all the design specifications and that it functions properly. Chips used in Printed Circuit Boards (PCBs) are packaged in Dual In-line Package (DIP), Pin Grid Array (PGA), Ball Grid Array (BGA), and Quad Flat Package (QFP).

**Transistors modeling:**

The transistor models are characterized by a figure of merit that depends on (a) performance, (b) level of integration and (c) cost. These are further influenced by a number of other factors including:

- Minimum feature size;
- Number of gates;
- Power dissipation;
- Die size;
- Gate delay;
- Testing;
- Reliability, and
- Production cost.

Therefore there are many issues that the designer needs to comprehend for a successful design.

## METAL-OXIDE-SEMICONDUCTOR (MOS) AND RELATED VLSI TECHNOLOGY

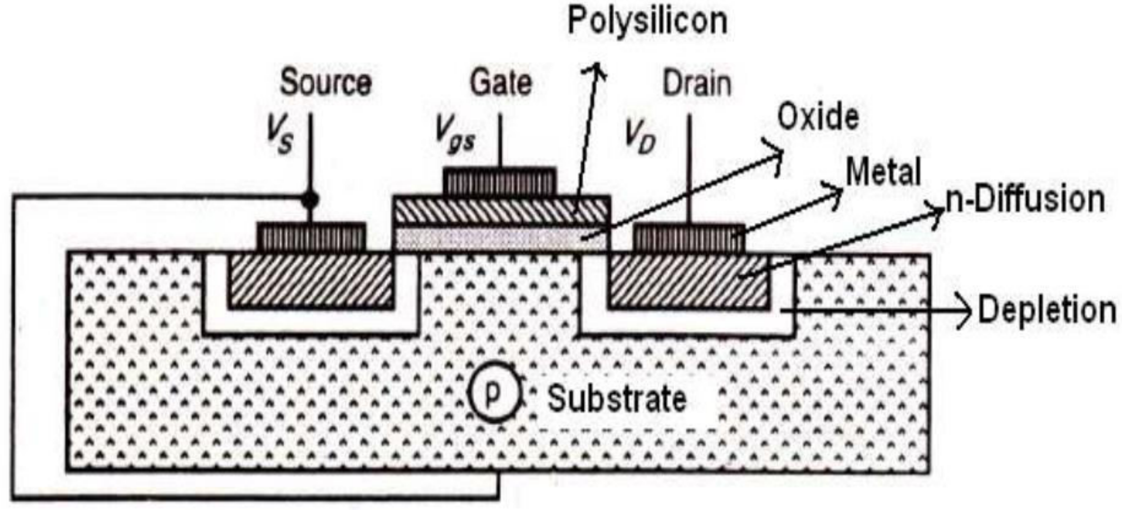
Within the bounds of MOS technology, the possible circuit realizations may be based on pMOS, nMOS, CMOS (which includes nMOS and pMOS transistors) and BiCMOS devices. Although CMOS is the dominant technology, some of the examples used to illustrate the design processes will be presented in nMOS form. The reasons for this are as follows:

- For nMOS technology, the design methodology and the design rules are easily learned, thus providing a simple but excellent introduction to structured design for VLSI.
- nMOS technology and design processes provide an excellent background for other technologies. In particular, some familiarity with nMOS allows a relatively easy transition to CMOS technology design.

Not only is VLSI technology providing the user with a new and more complex range of 'off the self' circuits, but VLSI design processes are such that system designers can readily design their own special circuit of considerable complexity. This provides a new degree of freedom for designers and it is probable that some very significant advances will results.

## BASIC MOS TRANSISTORS

**N-CHANNEL ENHANCEMENT TRANSISTOR:** nMOS devices are formed in a p-type substrate of moderate doping level. Two heavily doped n-type regions are created in substrate using doping process. They are called source and drain. A layer of $SiO_2$ is grown on the top surface of the substrate. The $SiO_2$ layer acts as an insulator. Next metal (poly silicon) is deposited on top of the $SiO_2$ and it works as gate electrode of the device. Next metal layer is grown at the top of source, drain, gate and substrate regions to provide different ohmic contacts. So it has four terminals.



From the device structure the p-type substrate forms pn-junction with the source(S) and drain(D) regions. Therefore the S and D are isolated from one another by these diodes. In normal operation these diodes are kept reverse biased at all times, since the drain will be at +ve voltage related to the source. The two pn-junctions can be effectively cut-off by simply connecting the substrate terminal to the source.

For an **enhancement-mode, n-channel MOSFET**, the four operational modes are:

***1. Cut-off , sub threshold or weak-inversion mode:***

**When $V_{gs} < V_t$:**

When Vgs< Vt, the transistor is turned off because two back to back diodes exits in series between D and S. These diodes prevent current conduction from D to S. So no current flows between D and S.

A more accurate model considers the effect of thermal energy on the Boltzmann distribution of electron energies which allow some of the more energetic electrons at the source to enter the channel and flow to the drain. This results in a subthreshold current that is an exponential function of gate to source voltage. While the current between drain and source should ideally be zero when the transistor is being used as a turned-off switch, there is a weak-inversion current, sometimes called subthreshold leakage.

2. ***$V_{gs} > V_t$ and $V_{ds}$ 0:***

A small +ve Vgs is applied on the gate terminal. Due to Vgs, holes in the P type layer close to the silicon dioxide layer under the gate to be repelled down into the P type substrate, and at the same time this positive potential on the gate attracts free electrons from the surrounding substrate material. These free electrons form a thin layer of charge carriers beneath the gate electrode (they can't reach the gate because of the insulating silicon dioxide layer) bridging the gap between the heavily doped source and drain areas. This layer is called channel and also sometimes called an "inversion layer" because applying the gate voltage has caused the P type material immediately under the gate to firstly become "intrinsic" and then an N type layer within the P type substrate.

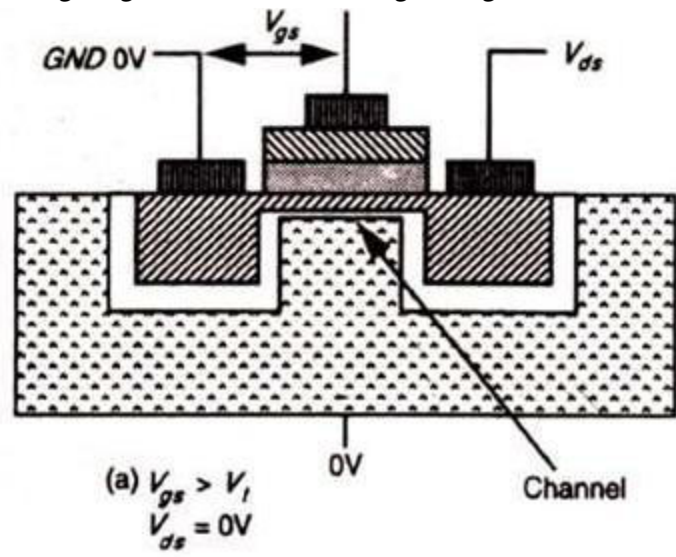
Any further increase in the gate voltage attracts more charge carriers into the inversion layer, so reducing its resistance, and increasing current flow between source and drain. Reducing the gate source voltage reduces current flow. When the power is switched off, the area beneath the gate reverts to P type once more. This method of operation is called "ENHANCEMENT MODE" as the application of gate source voltage makes a conducting channel "grow", therefore it enhances the channel. This MOSFET is called n-channel because the channel is populated with n-type carriers.

**Threshold voltage :** The gate voltage at which a sufficient no.of electrons accumulate under the gate region, to form a channel and start conduction between S and D is called the threshold voltage(Vt).For n-channel Vt should be +ve and for p-channel Vt will be –ve. Its value depends on the process of device fabrication.

**The gate and substrate form a parallel plate capacitor where $SiO_2$ acts as a dielectric.**

When we apply a positive voltage on its gate, the top plate of the capacitor will accumulate a positive charge. Similarly the bottom plate of the capacitor will accumulate a negative charge.

Due to this charge formation, it will develop an electrical field in vertical direction across the channel. It is the field which controls the amount of accumulated charge in the channel. So this voltage Vgs is called controlling voltage which determines the channel conductivity.

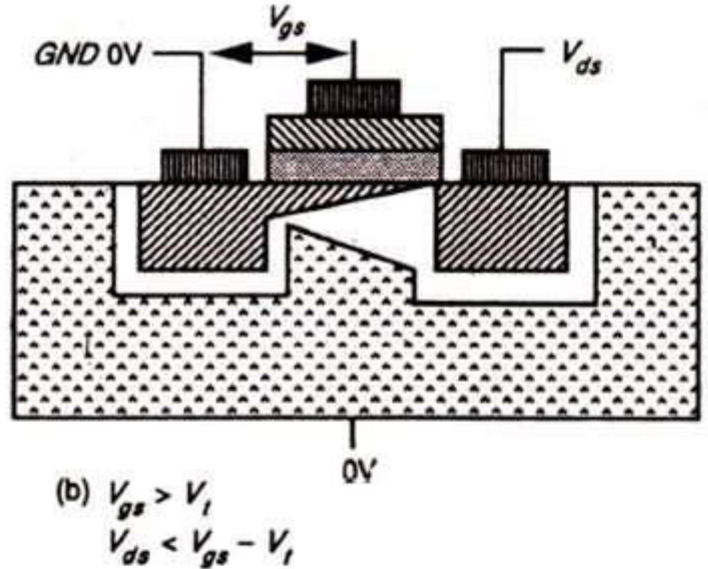



***Triode mode or linear region (also known as the ohmic mode***
***when $V_{gs} > V_t$ and $V_{ds} < ( V_{gs} – V_t )$***

When we apply a small amount of Vds on its drain, then the current will start flowing through the induced channel. The direction of current($I_D$) will be from D to S and the magnitude of $I_D$ depends on the density of electrons in the channel again which depends on Vgs.

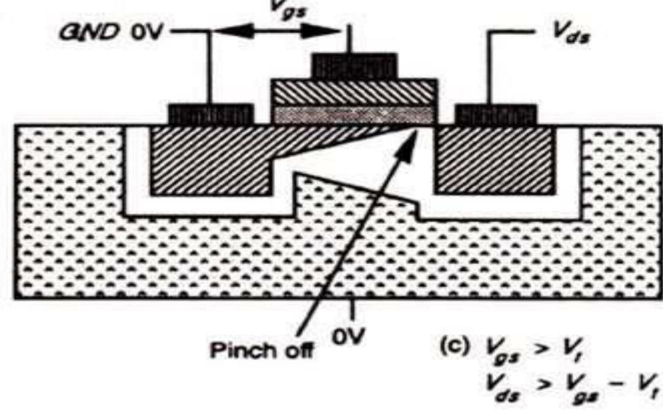
As Vds is increased, then current flows in the channel .There must be a corresponding IR drop = Vds along the channel. This develops a voltage between gate and channel varying with distance along the channel with the voltage being a maximum of Vgs at the source end. Due to this voltage variance across the channel, the channel is no longer uniform depth and its depth depends on the voltage across it. Therefore due to Vds, the channel shape will be tapered . The channel being deepst at the source end and shallowest at the drain end.

Since the effective gate voltage is Vg= Vgs - Vt (no current flows when Vgs < Vt), there will be voltage available to invert the channel at the drain end so long as Vds ≤ (Vgs - Vt )· The limiting condition comes when Vds= Vgs - Vt. For all voltages Vds < Vgs - Vt, the device operated in the non-saturated region.




***Saturation region when $V_{gs} > V_t$ and $V_{ds} > ( V_{gs} – V_t )$:***

Let us now consider the situation when Vds is increased to a level greater than Vgs - Vt. In this case, an IR drop equal to Vgs – Vt occurs over less than the whole length of the channel such that, near the drain, there is insufficient electric field available to give rise to an inversion layer to create the channel. The channel is, therefore, 'pinched off .Diffusion current completes the path from source to drain in this case, causing the channel to exhibit a high resistance and behave as a constant current source. This region, known as saturation, is characterized by almost constant current for increase of Vds above Vds = Vgs - Vt. In all cases, the channel will cease to exist and no current will flow when Vgs < Vt. Typically, for enhancement mode devices, Vt = 1 volt for VDD = 5 V or, in general terms, Vt = 0.2 VDD.

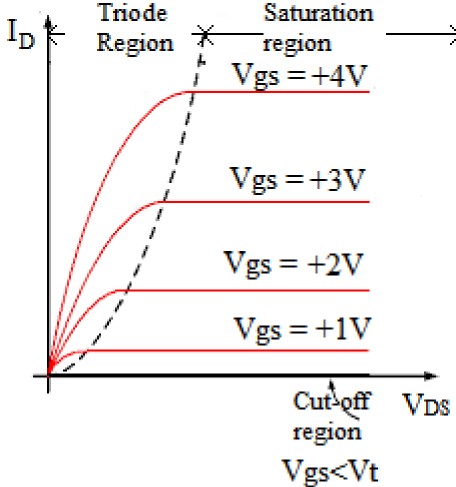


**$I_D$-$V_{DS}$ Characteristics of MOS Transistor :**

The graph below shows the $I_D$ Vs $V_{DS}$ characteristics of an n- MOS transistor for several values of Vgs .It is clear that there are two conduction states when the device is ON, they saturated state and the non-saturated state. The saturated curve is the flat portion and defines the saturation region. For Vgs < $V_{DS}$ + Vt, the nMOS device is conducting and $I_D$ is independent of $V_{DS}$.

For Vgs > $V_{DS}$ + Vth, the transistor is in the non-saturation region and the curve is a half parabola.

When the transistor is OFF (Vgs < Vt), then $I_D$ is zero for any $V_{DS}$ value.



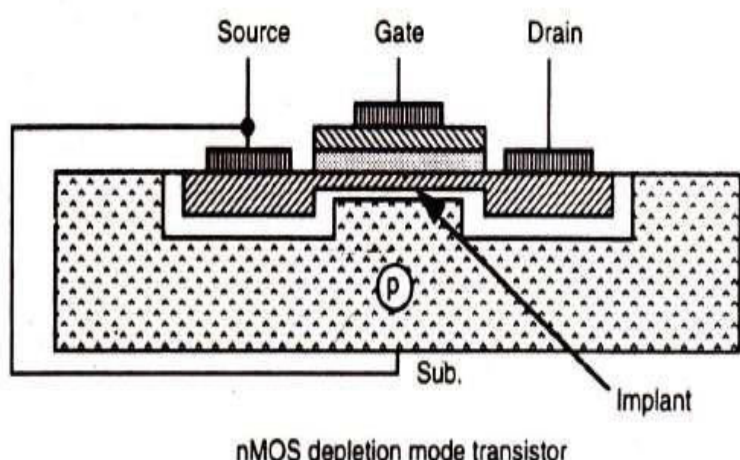

The boundary of the saturation/non-saturation bias states is a point seen for each curve in the graph as the intersection of the straight line of the saturated region with the quadratic curve of the non-saturated region. This intersection point occurs at the channel pinch off voltage called $V_{DSAT}$. $V_{DSAT}$ is defined as the minimum drain-source voltage that is required to keep the transistor in saturation for a given Vgs .

In the non-saturated state, the drain current initially increases almost linearly from the origin before bending in a parabolic response. Thus the name, ohmic or triode or linear for the non- saturated region. The drain current in saturation is virtually independent of $V_{DS}$ and the transistor acts as a current source. This is because there is no carrier inversion at the drain region of the channel. Carriers are pulled into the high electric field of the drain/substrate pn junction and ejected out of the drain terminal.

**N-CHANNEL DEPLETION MODE TRANSISTOR (DE-MOSFET):**

**Construction of a DEMOSFET:** Figure shows the construction of an N-channel depletion MOSFET. It consists of a highly doped P-type substrate into which two blocks of heavily doped N-type material are diffused forming the source and drain. An N-channel is formed by diffusion

between the source and drain. The type of impurity for the channel is the same as for the source and drain. Now a thin layer of $SiO_2$ dielectric is grown over the entire surface and holes are cut through the $SiO_2$(silicon-dioxide) layer to make contact with the N-type blocks (Source and Drain). Metal is deposited through the holes to provide drain and source terminals, and on the surface area between drain and source, a metal plate is deposited. This layer constitutes the gate. $Si0_2$ layer results in an extremely high input impedance of the order of $10^{10}$ to $10^{15}$ Q for this area. A P-channel DE-MOSFET is constructed like an N-channel DE-MOSFET, starting with an N-type substrate and diffusing P-type drain and source blocks and connecting them internally by a P-doped channel region.



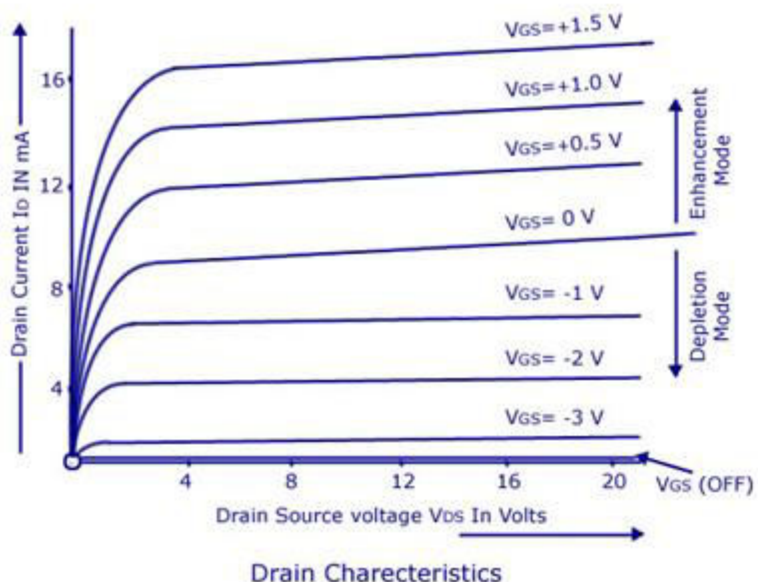
nMOS depletion mode transistor

**Operation of DEMOSFET:**

when the gate is made negative with respect to the substrate, the gate repels some of the negative charge carriers out of the N-channel. This creates a depletion region in the channel, aill and therefore, increases the channel resistance and reduces the drain urrent. The more negative the gate, the less the drain current. In this mode of operation the device is referred to as a *depletion-mode MOSFET.* Here too much negative gate voltage can pinch-off the channel.

On the other hand When the drain is made positive with respect to source, a drain current will flow, even with zero gate potential and the MOSFET is said to be operating in Enhancement mode. In this mode of operation gate attracts the negative charge carriers from the P-substrate to the N-channel and thus reduces the channel resistance and increases the drain-current. The more positive the gate is made, the more drain current flows.

So DE-MOSFET can be operated with either a positive or a negative gate. When gate is positive with respect to the source it operates in the enhancement mode and when the gate is negative with respect to the source, it operates in depletion-mode.



Drain Charecteristics

## IC PRODUCTION PROCESSES

**Integrated Circuit (IC)**

An Integrated Circuit (IC) is also called as chip or microchip. It is a semiconductor wafer in which millions of components are fabricated. The active and passive components such as resistors, diodes, transistors etc and external connections are usually fabricated in on extremely tiny single chip of silicon. All circuit components and interconnections are formed on single thin wafer (substrate) is called monolithic IC. IC is very small in size. It require microscope to see connections between components. The steps to fabricate IC chips is similar to the steps required to fabricate transistors, diodes etc. In IC chips, the fabrication of circuit elements such as transistors, diodes, capacitors etc. and their interconnections are done at same time. It has so many advantages such as extremely small size, small weight, low cost, low power consumption, high processing speed, easy replacement, etc. IC is the principal component in all electronic systems n. IC can function as amplifier, oscillator, timer, counter, computer memory etc.

The manufacturing of Integrated Circuits (IC) consists of following steps. The steps includes 8-20 patterned layers created into the substrate to form the complete integrated circuit.

IC production process steps:

Step1: Wafer preparation

Step2: Oxidation
Step3: Masking and lithography
Step4: Etching
Step5: Doping
Step6: Metallization
Step7: Testing
Step8: Packaging

## 1. Wafer Preparation:

The first step is wafer production. The wafer is a round slice of semiconductor material such as silicon. Silicon is preferred due to its characteristics. It is more suitable for manufacturing IC. It is the base or substrate for entire chip.

Wafer preparation requires three general processes which are SILICON REFINEMENT, CRYSTAL GROWTH and WAFER FORMATION.

**SILICON REFINEMENT**: Silicon is the most important semiconductor for the microelectronics industry. When compared to germanium, silicon excels for the following reasons:

(1) Si has a larger bandgap (1.1 eV for Si versus 0.66 eV for Ge).
(2) Si devices can operate at a higher temperature ($150^oC$ vs $100^oC$).
(3) Intrinsic resistivity is higher ($2.3 \times 10^5$ Ω-cm vs 47 Ω-cm).
(4) $SiO_2$ is more stable than $GeO_2$ which is also water soluble.
(5) Si is less costly.

Electronic-grade silicon (EGS), a polycrystalline material of high purity, is the starting material for the preparation of single crystal silicon. EGS is made from metallurgical-grade silicon (MGS) which in turn is made from quartzite, which is a relatively pure form of sand. MGS is purified by the following reaction:

$$Si\ (solid) + 3HCl\ (gas) \rightarrow SiHCl_3\ (gas) + H_2\ (gas) + heat$$

The boiling point of trichlorosilane ($SiHCl_3$) is $32^oC$ and can be readily purified using fractional distillation. EGS is formed by reacting trichlorosilane with hydrogen:

$$2SiHCl_3\ (gas) + 2H_2\ (gas) \rightarrow 2Si\ (solid) + 6HCl\ (gas)$$

Electronic-grade silicon is the raw material used to prepare device. This is called single crystal silicon.

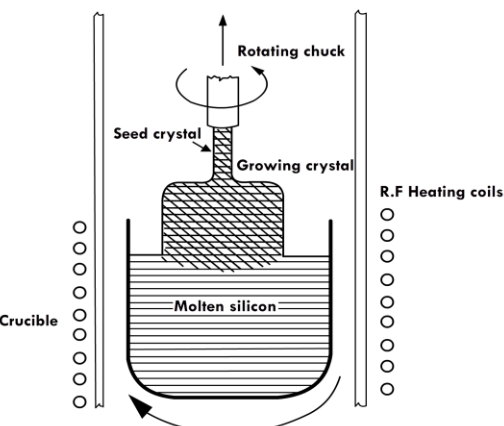
**CRYSTAL GROWTH:** There are two main techniques for converting polycrystalline EGS into a single crystal ingot, which are used to obtain the final wafers.

1. Czochralski technique (CZ) - this is the dominant technique for manufacturing single crystals. It is especially suited for the large wafers that are currently used in IC fabrication.
2. Float zone technique - this is mainly used for small sized wafers. The float zone technique is used for producing specialty wafers that have low oxygen impurity concentration.

**Czochralski technique(CZ):**
A schematic of this growth process is shown in figure. The various components of the process are

1. Furnace
2. Crystal pulling mechanism
3. Ambient control - atmosphere
4. Control system

The starting material for the CZ process is electronic grade silicon, which is melted in the furnace. To minimize contamination, the crucible is made of quartz .



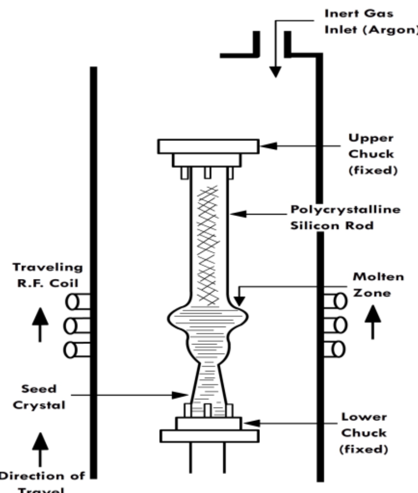

The furnace is heated above $1500^o$ C, since Si melting point is $1412^o$ C. A precisely oriented rod-mounted seed crystal is dipped into the molten Silicon. The seed crystal's rod is slowly pulled upwards and rotated simultaneously. The furnace is rotated in the direction opposite to the crystal puller. The molten Si sticks to the seed crystal and starts to solidify with the same orientation as the seed crystal is withdrawn. Thus, a single crystal ingot is obtained.

To create doped crystals, the dopant material is added to the Si melt so that it can be incorporated in the growing crystal. By precisely controlling the temperature gradients, speed of pulling and speed of rotation of the crystal puller, it is possible to extract a large, single-crystal cylindrical ingot from the melt. This ingot is further processed to get the wafers that are used for fabrication.

Fig: Single crystal Si ingot

**Float zone technique:**

The float zone technique is suited for small wafer production, with low oxygen impurity. The schematic of the process is shown in figure . A polycrystalline EGS rod is fused with the single crystal seed of desired orientation. This is taken in an inert gas furnace and then melted along the length of the rod by a traveling radio frequency (RF) coil. The RF coil starts from the fused region, containing the seed, and travels up, as shown in figure . When the molten region solidifies, it has the same orientation as the seed. The furnace is filled with an inert gas like argon to reduce gaseous impurities.




Also, since no crucible is needed it can be used to produce oxygen 'free' Si wafers. The difficulty is to extend this technique for large wafers, since the process produces large number of dislocations. It is used for small specialty applications requiring low oxygen content wafers.

**WAFER FORMATION:**

After the single crystal is obtained, this needs to be further processed to produce the wafers. For this, the wafers need to be shaped and cut. Usually, industrial grade diamond tipped saws are used for this process. The shaping operations consist of two steps

1. The seed and tang ends of the ingot are removed.
2. The surface of the ingot is ground to get a uniform diameter across the length of the ingot.

Before further processing, the ingots are checked for resistivity and orientation. Resistivity is checked by a four point probe technique and can be used to confirm the dopant concentration. This is usually done along the length of the ingot to ensure uniformity. Orientation is measured by x-ray diffraction at the ends (after grinding).

After the orientation and resistivity checks, one or more flats are ground along the length of the ingot. After making the flats, the individual wafers are sliced per the required thickness. After cutting, the wafers are chemically etched to remove any damaged and contaminated regions. This is usually done in an acid bath with a mixture of hydrofluoric acid, nitric acid, and acetic acid. After etching, the surfaces are polished, first a rough abrasive polish, followed by a

chemical mechanical polishing (CMP) procedure. In CMP, a slurry of fine SiO2 particles suspended in aqueous NaOH solution is used. The pad is usually a polyester material. Polishing happens both due to mechanical abrasion and also reaction of the silicon with the NaOH solution.

Wafers are typically single side or double side polished. Large wafers are usually double side polished so that the backside of the wafers can be used for patterning. But wafer handling for double side polished wafers should be carefully controlled to avoid scratches on the backside. Typical 300 mm wafers used for IC manufacture are handled by robot arms and these are made of ceramics to minimize scratches. Smaller wafers (3" and 4" wafers) used in labs are usually single side polished. After polishing, the wafers are subjected to a final inspection before they are packed and shipped to the fab.

## 2.Oxidation:

Oxidation is the process in which oxygen (dry oxidation) or $H_2O$(wet oxidation) molecules convert silicon layers on top of the wafer to silicon dioxide. The chemical reaction of silicon and oxygen already starts at room temperature but stops after a very thin native oxide film. For an effective oxidation rate the wafer must be settled to a furnace with oxygen or water vapor at elevated temperatures.

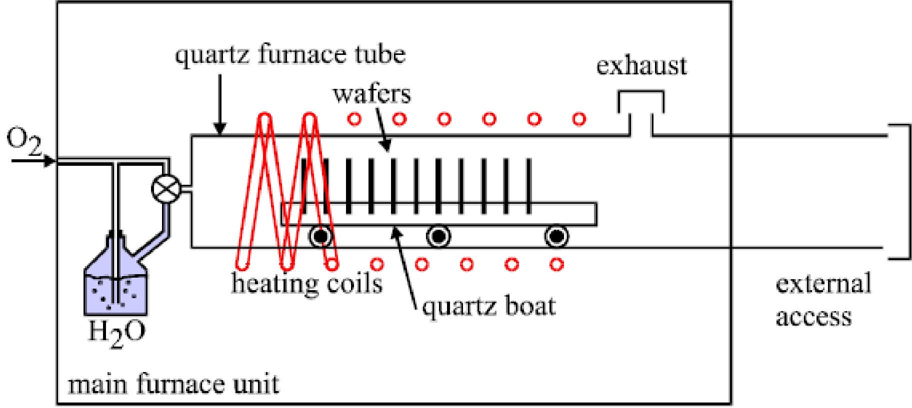
Purpose of oxide layers: They can be

- Part of the active device
- Used as masks to protect against diffusion or ion implantation
- Used as protecting layer at the end of device fabrication

Silicon dioxide layers are used as high-quality insulators or masks for ion implantation. The ability of silicon to form high quality silicon dioxide is an important reason, why silicon is still the dominating material in IC fabrication.

**Thermal oxidation** is a way to produce a thin layer of $SiO_2$on the surface of a substrate. The thermal oxidation of $SiO_2$ consists of exposing the Si substrate to an oxidation environment of $O_2$ or $H_2O$ at elevated temperature. Thermal oxidation is accomplished by using an oxidation furnace which provides the heat needed to elevate the oxidizing ambient temperature.

The heating system usually consists of several heating coils that control temperature around the furnace tube. The wafers are placed in quartz glass ware called boat. The boat can contain many wafers typically 50 or more. The oxidizing agent(oxygen or steam) then enters the process tube through its source end, subsequently diffusing to the wafers where oxidation occurs.




**Oxidation methods:** Two types of oxidation methods are there

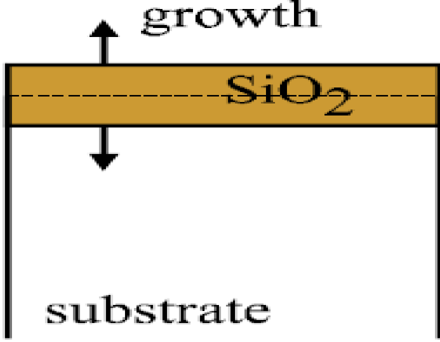
1. Wet oxidation
2. Dry oxidation

1.**Wet oxidation** : During wet oxidation, the silicon wafer is placed into an atmosphere of water vapor ($H_2O$) and the ensuing chemical reaction is between the water vapor molecules and the

solid silicon atoms (Si) on the surface of the wafer, with hydrogen gas (H ) released as a byproduct.

$$Si(s)+ 2H_2O(l) \rightarrow SiO_2(s)+2H_2(gas)$$

These oxidation reactions occur at the Si – $SiO_2$ interface. As the oxide grows, the Si – $SiO_2$ interface will always be below the original Si wafer surface. The $SiO_2$ surface on the other hand, is always above the original Si surface. so oxide layer grows in both directions from the original substrate surface (approx. 50/50)




It is evident that wet oxidation operates with much higher oxidation rates than dry oxidation, up to approximately 600nm/h. The reason is the ability of hydroxide ($OH^-$) to diffuse through the already-grown oxide much quicker than $O_2$, effectively widening the oxidation rate bottleneck when growing thick oxides, which is the diffusion of species. Due to the fast growth rate, wet oxidation is generally used where thick oxides are required, such as insulation and passivation layers, masking layers, and for blanket field oxides.

**2.Dry oxidation:** During dry oxidation, the Si wafer react with the ambient oxygen, forming a layer of $SiO_2$ on its surface.

$$Si(s) + O_2(vapor) \rightarrow SiO_2(s)$$

The oxide films resulting from a dry oxidation process have a better quality than those grown in a wet environment, which makes them more desirable when high quality oxides are needed. Dry oxidation is generally used to grow films not thicker than 100nm or as a second step in the growth of thicker films, after wet oxidation has already been used to obtain a desired thickness. The application of a second step is only meant to improve the quality of the thick oxide.

## 3. Masking and lithography:

**Lithography:** An IC consists of many microscopic regions(implantation regions and contact windows)on the wafer surface that make up the devices and interconnections as per the circuit. In the planner process, the regions are fabricated by steps that add, alter or remove in selected areas of the wafer surface. Each layer is determined by geometric pattern representing circuit design information.

Lithography is a process of drawing patterns on a silicon wafer. Different lithographic techniques are available which are **photolithography, Electron lithography, X-ray lithography** and **Ion lithography.**

**Photolithography**:

To protect some area of wafer when working on another area, a process called **photolithography** is used. The process of photolithography includes masking with a photographic mask and photo etching.

Photolithography is the transfer of an image using photographic techniques. It Uses light radiation to expose a coating of photoresist on the surface of the wafer. Common light source used in wafer processing is UV light due to its short wave length.

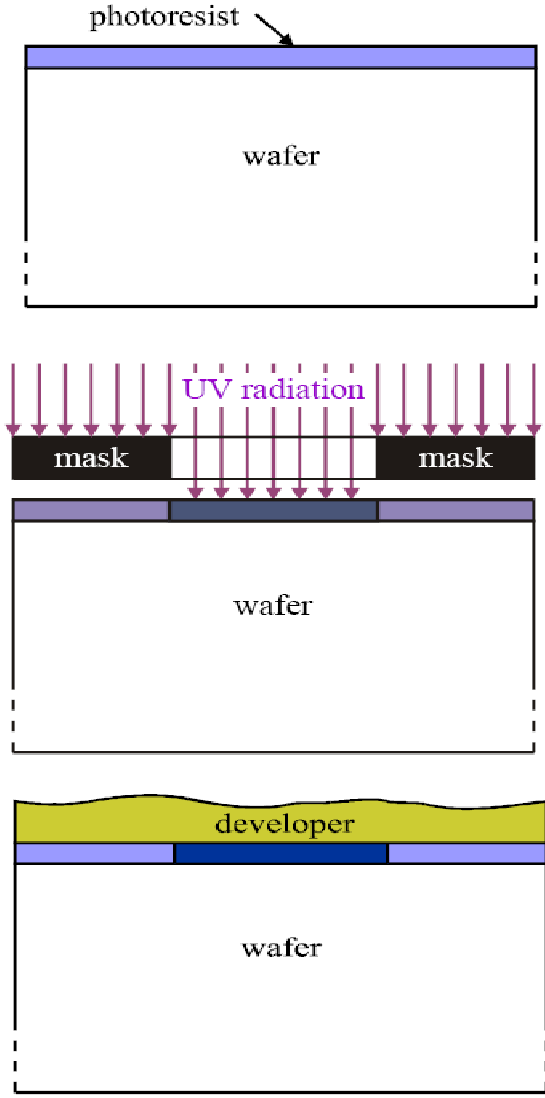
Photolithography transfers designer generated information (device placement and interconnections) to an actual IC structure using masks which contain the geometrical information. The process of photolithography is repeated many times in manufacture of an IC to build up device structures and interconnections.

**Photo mask**:

It is important component in photolithography. It contains blue print of the designed circuit. Using the photo mask, specific images of detailed devices are transferred on to the surface of the silicon wafer. A single photo mask plate produces identical images on 1000's of wafers. The quality of the photo mask determines the quality of semiconductor chips. The material used for building photo masks is quartz plate upon which detailed images or patterns are formed. The patterns are then transferred on to the wafer surface by exposing light through the quartz plate.

Each mask contains only layer of the circuit. A set of masks, each defining one pattern layer, is fed into a photolithography machine and individually selected for exposure to form the desired pattern on the wafer. Circuit elements such as transistors, capacitors and resistors are created by those patterns of many layers.

Photolithography process: First step in photolithography is to coat the surface with approx 1 μm of photoresist(PR). Photoresist is an organic polymer i.e sensitive to light radiation in a certain wavelength range. The sensitivity causes either an increase or decrease in solubility of the polymer to certain chemicals. The PR is then exposed to UV (ultraviolet) radiation through a mask. The masks generated from information about device placement and connection. The UV radiation causes a chemical change in the PR. The PR is then developed using a chemical developer.



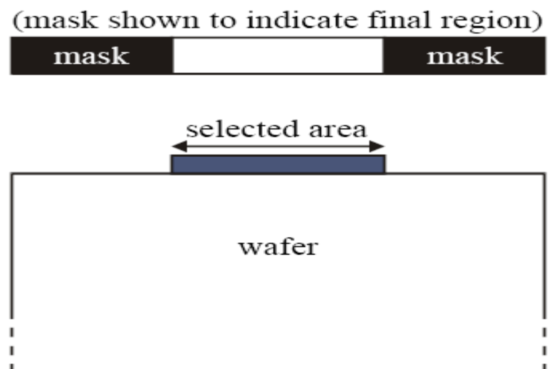

There are two types of PR. 1. Negative PR 2. Positive PR

**Negative PR:** A negative PR is hardened against the developer by the UV radiation, and hence remains on the surface where UV shone through the mask.

**Positive PR:** A positive PR is the opposite, it is removed where the UV shone through the mask

**EXAMPLE: Negative PR**

Assume a negative PR for this example, so the PR on the sides will be weakened and removed by the developer. Once the developer has been washed off, the result is PR in the region corresponding to the transparent part of the mask. Subsequent processing steps will use this structure to form device areas, interconnects, etc.




## 4. Etching:

**Etching** is the process of using strong acid or etchant to cut into the unprotected parts of a metal surface to create a design. It removes material selectively from the surface of wafer to create patterns. The pattern is defined by etching mask. The parts of material are protected by this etching mask. Etching is after lithography.

Etching is of two types:

1. wet etching　　　　2.Dry etching

**wet etching:**

Wet etching uses an acid, to remove a target material. Etchant is selected to chemically attack the specific material to be removed and not the protective layer. For silicon, the most commonly used etchants are mixtures of nitric acid and hydrofluoric acid in water or acetic acid. Wet etching is good and fairly cheap and capable of processing many wafers quickly. The disadvantage is that wet etching does not allow the smaller critical geometries that are needed for today chips.

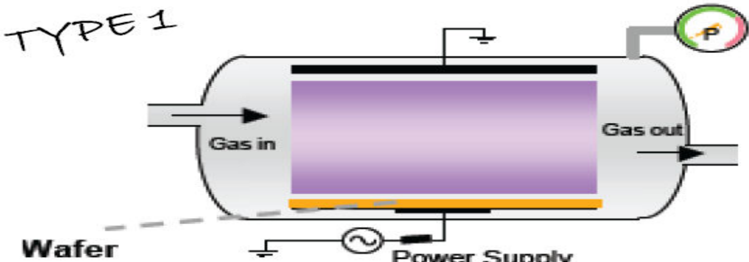
**Dry etching:**

Dry etching uses gas instead of chemical etchants. It is capable of producing critical geometries that are very small. Example: Plasma etching

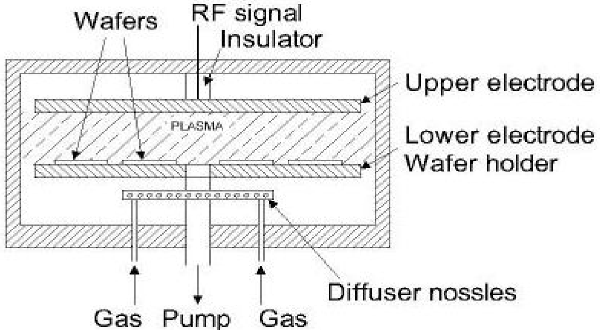
Plasma etching: Plasma etching uses a gas that is subjected to an intense electric field to generate the plasma state( Plasma is an ionized gas composed of equal no.of positive and negative charges and a different no.of un ionizes molecules). The electric field is produced with coils that are wrapped around the chamber and exposed to a high level RF source.

There are two different versions of this type of etching based on the shape of the chamber used.

1. One consists of a barrel type chamber where the wafers are placed sitting up while the gas is flowed over the wafers and out through an exhaust pipe.




2. The second type uses a parallel plate reactor . Here there are two plates that are used to give the gas the electric field rather than the coil that is wrapped around the barrel chamber.




In plasma form, the gases used are very reactive, providing effective etching of the exposed surface. Plasma etching provides good critical geometry but the wafer can be damaged from the RF radiation.

# ETCHANT and ETCHED LAYER

| Material to be etched | Wet etchants | Dry / Plasma etchants |
|---|---|---|
| Silicon (Si) | Nitric acid ($HNO_3$) + hydrofluoric acid (HF)[3] | •$CF_4$, $SF_6$, $NF_3$[4]<br>•$Cl_2$, $CCl_2F_2$[4] |
| Silicon dioxide ($SiO_2$) | •Hydrofluoric acid (HF)[3]<br>•Buffered oxide etch [BOE]: ammonium fluoride ($NH_4F$) and hydrofluoric acid (HF)[3] | $CF_4$, $SF_6$, $NF_3$[4] |
| Photoresist | Piranha etch: sulfuric acid ($H_2SO_4$) + hydrogen peroxide ($H_2O_2$) | $O_2$ (ashing) |
| Aluminium (Al) | 80% phosphoric acid ($H_3PO_4$) + 5% acetic acid<br>+ 5% nitric acid ($HNO_3$) + 10% water ($H_2O$) at 35–45 °C[3] | $Cl_2$, $CCl_4$, $SiCl_4$, $BCl_3$[4] |

**Types of etching profiles:** The shape of the feature that is etched is called the etch profile. There are two types of etch profiles.

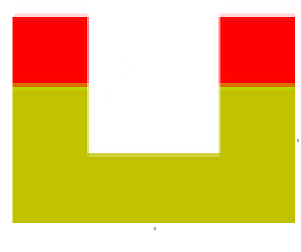
1. Isotropic 2. Anisotropic

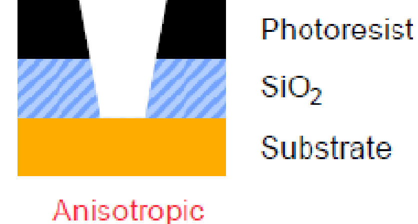
To perform **etching** in all directions at same time, isotropic etching will be used. Anisotropic etching is faster in one direction.

1. Isotropic etch profile: Etched equally in all directions. Wet etches gives the isotropic etch profile. Some dry etches also give the isotropic etch profile. A perfectly isotropic etch produces round side walls.

**Isotropic**

1. 2.Anisotropic etch profile: Etched in a preferred direction only. Dry etches gives the anisotropic etch profile. Anisotropic profiles are needed to transfer lithographic patterns for small features. A perfectly anisotropic etch produces vertical sidewalls.






**Anisotropic**

## Wet and Dry Etching

| | Wet | Dry |
|---|---|---|
| Method | Chemical Solutions | Ion Bombardment or Chemical Reactive |
| Environment and Equipment | Atmosphere, Bath | Vacuum Chamber |
| Advantage | 1) Low cost, easy to implement<br>2) High etching rate<br>3) Good selectivity for most materials | 1) Capable of defining small feature size (< 100 nm) |
| Disadvantage | 1) Inadequate for defining feature size < 1um<br>2) Potential of chemical handling hazards<br>3) Wafer contamination issues | 1) High cost, hard to implement<br>2) low throughput<br>3) Poor selectivity<br>4) Potential radiation damage |
| Directionality | Isotropic<br>(Except for etching Crystalline Materials) | Anisotropic |

# 5.Doping:

Inorder to fabricate semiconductor devices, a controlled amount of impurities are added selectively into the single crystal wafers. Three methods are used for controlled doping of a semiconductor. They are

1. Epitaxy  2. Diffusion  3. Ion implantation

**1. Epitaxy :**

In this process a thin layer of single crystal semiconductor (nm to um) is grown on an already existing crystalline substrate such that the grown film has same lattice as the substrate. There are two types of epitaxy. a. Homo epitaxy  b. Hetero epitaxy

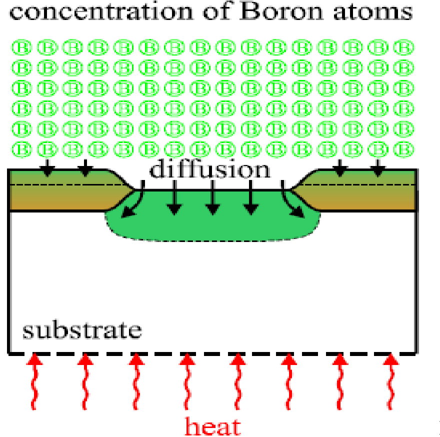
a. **Homo epitaxy:** In which same layer is grown over the substrate.
Example: Si is growing on Si substrate.

b. **Hetero epitaxy:** In which different layer is grown over the substrate.
Example: AlGaAs is growing on GaAs.

**2. Diffusion**:

By using epitaxy we can grow a layer with controlled doping but we can't control the doping of selective regions of the semiconductor surface. It means that epitaxial growth takes place throughout the surface i.e it is non-selective. Inorder to get selective doping, the most commonly used technique is diffusion.

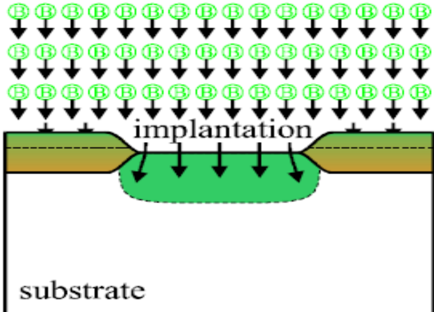
In this method p and n regions are created by adding dopants into the wafer. The wafers are placed in an oven which is made up of quartz and it is surrounded with heating elements. Then the wafers are heated at a temperature of about 1500-2200°F. The inert gas carries the dopant chemical. The dopant and gas is passed through the wafers and finally the dopant will get deposited on the wafer. This method can only be used for large areas. For small areas it will be difficult and it may not be accurate.




m

**2.Ion implantation**:

This is also a method used for adding dopants. In this method, dopant gas such as phosphine or boron trichloride will be ionized first. Then it provides a beam of high energy dopant ions to the specified regions of wafer. It will penetrate the wafer. The depth of the penetration depends on the energy of the beam. By altering the beam energy, it is possible to control the depth of penetration of dopants into the wafer. The beam current and time of exposure is used to control the amount of dopant. This method is slower than atomic diffusion process. First it points the wafer that where it is needed and shoot the dopants to the place where it is required.




## 6. Metallization:

Metallization is a process of adding a layer of metal on the surface of wafer.

Functions of conductive materials on wafer surface:

- used to create contact with silicon
- form certain components(e.g gates) of IC devices
- provide interconnecting conduction paths between devices on chip
- connect the chip to external circuits

Metallization materials:

Aluminium: A thin layer of aluminum is deposited over the whole wafer. Aluminium is selected because it is a good conductor, has good mechanical bond with silicon, forms low resistance contact and it can be applied and patterned with single deposition and etching process.

Other materials: poly silicon, gold, silicides and nitrides.

## 7. Testing:

After the wafer has been processed and the final metallization pattern defined, it is placed in a holder under a microscope and is aligned for testing by a multiple-point probe .The probe contacts the various pads on an individual circuit and a series of tests are made of the electrical properties of the device. The various tests are conducted automatically in a very short time ranging from a few milliseconds for a simple circuit to 30 seconds or more for a complex chip. The test results are fed into a computer, and a decision is made regarding the acceptability of the circuit. If the chip is defective or the circuit falls below specifications, the computer instructs the test probe to mark the circuit with a dot of ink. The probe automatically steps the prescribed distance to the next chip on the wafer and repeats the process. After all of the circuits have been tested and substandard ones marked, the wafer is removed from the testing machine, scribed between the circuits, and broken apart .In the testing process, information from tests on each circuit can be printed out to facilitate analysis of the rejected ones or to evaluate the fabrication process for possible modification.

## 8.Packaging:

Packaging is used to connect the IC to the outside world.

Functions of packaging:

- Packages protect the IC from damaging external influences like Moisture, Dust, Vibration, Shock, Lightning, Magnets, etc.

- The chip is attached to a lead frame and encapsulated inside a package. Lead frame allows electrical signals to be sent and received to and from semiconductor devices.
- Packages effectively release the heat generated by the chip during its operation.
- Packages allow for enlargement of terminals size that makes the chips much easier to handle.

IC packages are classified according to the way they are mounted on the PCB as either pin through hole mounted or surface mounted.

**Pin- through-hole package:** Pin through hole packages have pins(leads) that are inserted through holes in the PCB and can be soldered to conductors on the opposite side.
**Surface mount technology(SMT):**pins of surface mounted packages are soldered directly to conductors on one side of the board, leaving other side free for additional circuits.

IC packages can be further grouped into three general categories; Dual In-line Packages, Chip Carriers and Grid Arrays. All the packages, regardless of the category has a body style that scales with pin count. That is the name of the package does not determine the physical size of the package, the number of pins do.
**1.** Dual In-line Packages [DIP], or Dual In-Line [DIL] packages are packages with two rows of leads on two sides of the package. DIP ICs may be through-hole [PDIP or CERDIP] or SMT package [SOJ or SOIC].
**2.** Quad Flat Packs or Chip Carriers are square packages [or nearly square], with leads on all four sides . Chip Carriers, as in PLCCs and other variants are strictly Surface Mount Technology (SMT).

**3.** Grid Arrays are those type packages that have their pins arranged in a grid.
The pin grid may consist of Leads, pads, or solder balls on an area array. The through hole variant is called a PGA, while the SMT variant might be called LGA or BGA.

**MOS AND CMOS FABRICATION PROCESS**

## nMOS fabrication process:

nMOS FABRICATION: Using the basic processes mentioned in the previous section, typical processing steps of the poly-silicon gate self-aligning nMOS technology are given below. The fabrication of nMOS can be considered a standard process. The advantages of this process over the other processes are that it is conceptually and physically simpler than other processes because it requires less photolithography steps. It has high functional density, good speed.

The major drawback of nMOS process is its high absolute power consumption and its electrical asymmetry.CMOS is replacing nMOS as the standard process because it minimizes both of the above disadvantages. But fabrication process used for nMOS is relevant to CMOS and BiCMOS, This may be viewed as involving additional fabrication steps.
Figure shows the step-by-step production of the transistor.

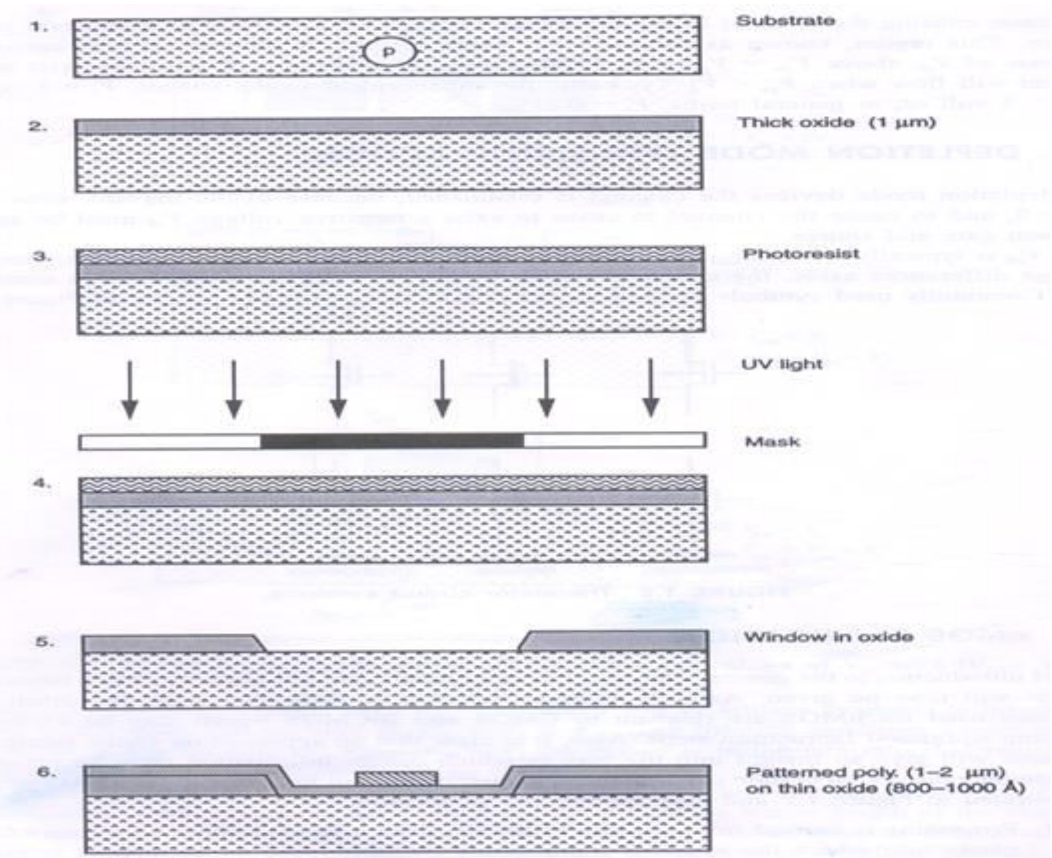
1.Processing is carried out on a thin wafer cut from a single crystal of silicon of high purity into which the required p-impurities are introduced as the crystal is grown. Such wafers are typically 75 to 150 mm in diameter and 0.4 mm thick and are doped with, say, boron to impurity concentrations of $10^{15}/cm^3$ to $10^{16}/cm^3$, giving resistivity in the approximate range 25 ohm cm to 2 ohm cm.

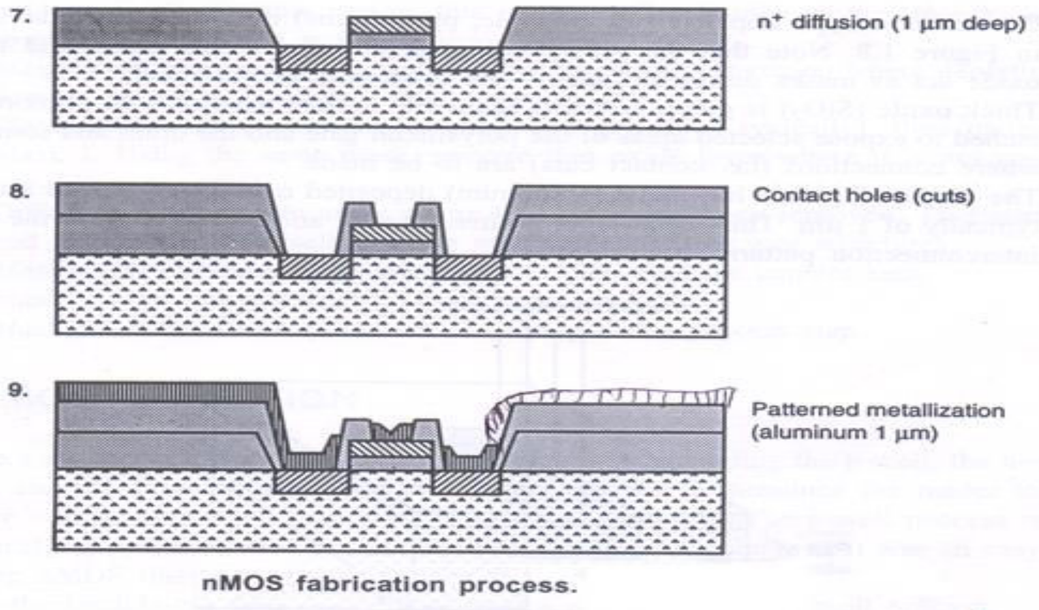
2.The next step is to grow a thick silicon dioxide ($SiO_2$) layer, typically of 1 μm thickness all over the wafer surface to protect the surface. This oxide layer will act as a barrier to dopant during subsequent processing and provide an insulting layer on which other patterned layers can be formed.

3. The surface is now covered with a photoresist which is deposited onto the wafer and to achieve an even distribution of the required thickness.

4.The photoresist layer is then exposed to ultraviolet light through a mask which defines those regions into which diffusion is to take place together with transistor channels. Assume, for example, that those areas exposed to ultraviolet radiation are polymerized (hardened), but that the areas required for diffusion are shielded by the mask and remain unaffected.

5. These areas are subsequently readily etched away together with the underlying silicon dioxide so that the wafer surface is exposed in the window defined by the mask.






nMOS fabrication process.

6. The remaining photoresist is removed and a thin layer of $SiO_2$ (0.1 $\mu m$ typical) is grown over the entire chip surface and then polysilicon is deposited on top of this to form the gate structure. The polysilicon layer consists of heavily doped polysilicon deposited by chemical vapor deposition (CVD). In the fabrication of fine pattern devices, precise control of thickness, impurity concentration, and resistivity is necessary.

7. Further photoresist coating and masking allows the polysilicon to be patterned (as shown in Step 6), and then the thin oxide is removed to expose areas into which n-type impurities are to be diffused to form the source and drain as shown. Diffusion is achieved by heating the wafer to a high temperature and passing a gas containing the desired n-type impurity (for example, phosphorus) over the surface as indicated .

8. Thick oxide ($SiO_2$) is grown over all again and is then masked with photoresist and etched to expose selected areas of the polysilicon gate and the drain and source areas where connections (i.e. contact cuts) are to be made.

9. The whole chip then has metal (aluminum) deposited over its surface to a thickness typically of 1µm.This metal layer is then masked and etched to form the required interconnection pattern.

It will be seen that the process revolves around the formation or deposition and patterning of three layers, separated by silicon dioxide insulation. The layers are diffused within the substrate, polysilicon on oxide on the substrate, and metal insulated again by oxide.

To form depletion mode devices it is only necessary to introduce a masked ion implantation step between Steps 5 and 6 in Figure. Again, the thick oxide acts as a mask and this process stage is also self-aligning.

**CMOS FABRICATION PROCESS**:

CMOS fabrication can be accomplished using either of the three technologies:

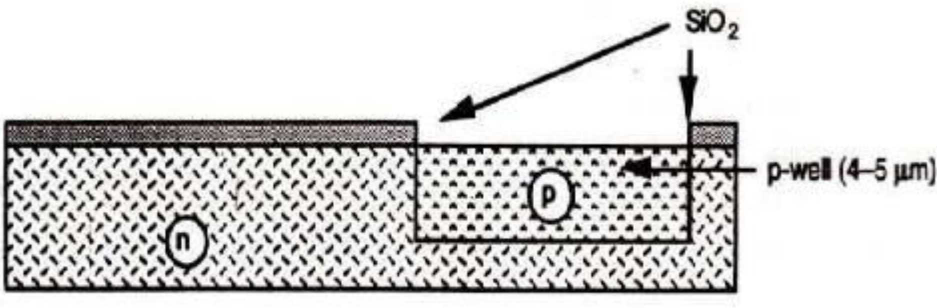
- N-well/P-well technologies
- Twin well technology
- Silicon On Insulator (SOI)

Among these methods the p-well process is widely used in practice and the n-well process is also popular, particularly as it is an easy retrofit to existing nMOS lines

**The P-well Process**

The basic processing steps used for P-Well Process are of the same as those used for nMOS fabrication.

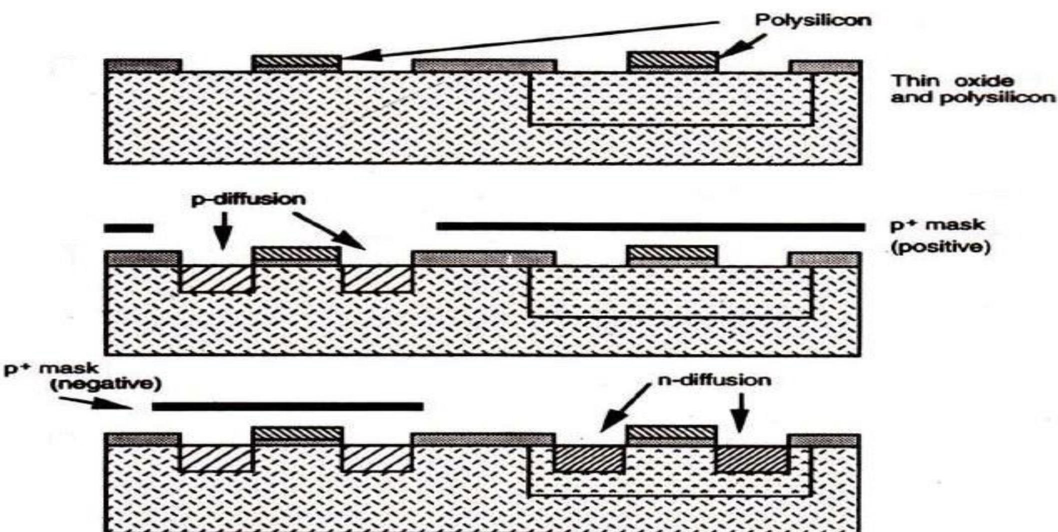
The P-Well structure consists of an n-type substrate in which p-devices may be formed by suitable masking and diffusion and, in order to accommodate n-type devices, a deep p-well is diffused into the n-type substrate as shown in the figure below



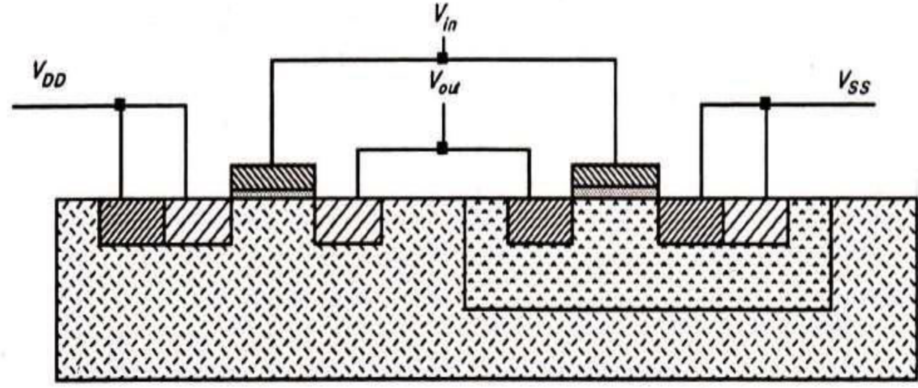

.

This diffusion must be carried out with special care since the p-well doping concentration and depth will affect the threshold voltages as well as the breakdown voltages of the n-transistors. To achieve low threshold voltages (0.6 to 1.0 V), we need either deep well diffusion or high well resistivity. However, deep wells require larger spacing between the n- and p-type transistors and wires because of lateral diffusion and therefore a larger chip area. The p-wells act as substrates for the n-devices within the parent n-substrate, and, provided that voltage polarity restrictions are observed, the two areas are electrically isolated.

In all other respects- like masking, patterning, and diffusion-the process is similar to nMOS fabrication.




However, since there are now in effect two substrates, two substrate connections *($V_{DD}$* and Vss) are required. The diagram below shows the CMOS p-well inverter showing VDD and Vss substrate connections




.

In summary, typical processing steps are:

- *Mask* 1-defines the areas in which the deep p-well diffusions are to take place.
- *Mask* 2-defines the thinox regions, namely those areas where the thick oxide is to be stripped and thin oxide grown to accommodate p- and n-transistors and diffusion Wires.
- *Mask* 3-used to pattern the polysilicon layer which is deposited after the thin oxide.
- *Mask* 4-A p-plus mask is now used (to be in effect 'Anded' with Mask 2) to define all areas where p-diffusion is to take place.
- *Mask* 5- This is usually performed using the negative form of the p-plus mask and, with Mask 2, defines those areas where n-type diffusion is to take place.

- *Mask* 6-Contact cuts are now defined.
- *Mask* 7- The metal layer pattern is defined by this mask.
- *Mask* 8-An overall passivation (overglass) layer is now applied and Mask 8 is needed to define the openings for access to bonding pads.

**The N-well Process:**

N-Well CMOS fabrication requires that both n-channel and p-channel transistors be built on the same chip substrate. To accommodate this, special regions are created with a semiconductor type opposite to the substrate type. The regions thus formed are called wells or tubs. In an n-type substrate, we can create a p-well or alternatively, an n-well is created in a p-type substrate. We present here a simple n-well CMOS fabrication process, in which the NMOS transistor is created in the p-type substrate, and the PMOS in the n-well, which is built-in into the p-type substrate.

Historically, fabrication started with p-well technology but now it has been completely shifted to n-well technology. The main reason for this is that, "n-well sheet resistance can be made lower than p-well sheet resistance" (electrons are more mobile than holes), lower substrate bias effects on transistor threshold voltage and inherently lower parasitic capacitances associated with source and drain regions.

The simplified process sequence for the fabrication of CMOS integrated circuits on a p-type silicon substrate is as follows:

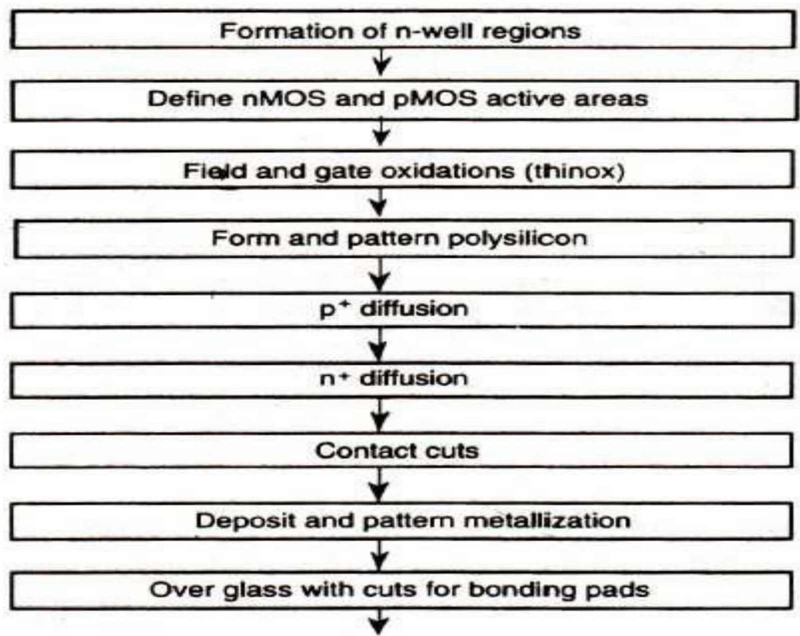
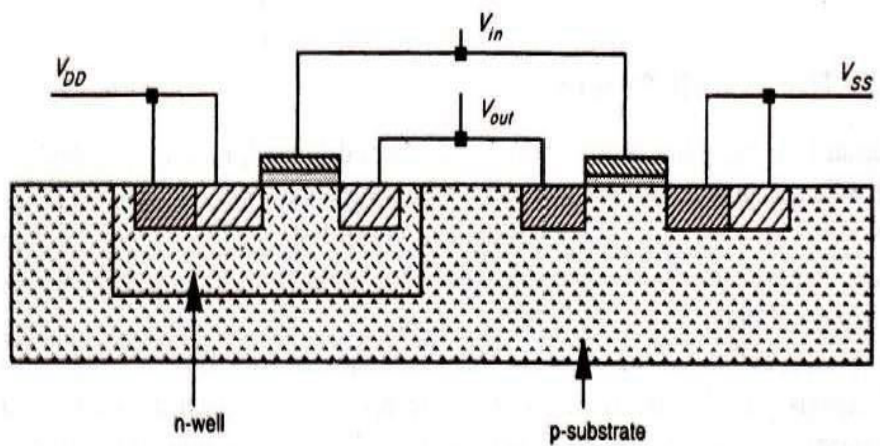



**Fig.n-well fabrication steps**

The first mask defines the n-well regions. This is followed by a low dose phosphorus implant driven in by a high temperature diffusion step to form the n-wells. The well depth is optimized to ensure against p-substrate to $p^+$ diffusion breakdown without compromising the n-well to $n^+$ mask separation. The next steps are to define the devices and diffusion paths, grow field oxide, deposit and pattern the polysilicon, carry out the diffusions, make contact cuts, and finally metalize as before.

It will be seen that an n+ mask and its complement may be used to define the n- and p-diffusion regions respectively. These same masks also include the $V_{DD}$ and $V_{SS}$ contacts (respectively). It should be noted that, alternatively, we could have used a $p^+$ mask and its complement, since the $n^+$ and $p^+$ masks are generally complementary.

The below Figure will shows an inverter circuit fabricated by the n-well process.




Due to differences in charge carrier motilities, the n-well process creates non-optimum p-channel characteristics. However, in many CMOS designs (such as domino-logic and dynamic logic structures), this is relatively unimportant since they contain a preponderance of n-channel devices. Thus the n-channel transistors are mainly those used to form logic elements, providing speed and high density of elements.

Latch-up problems can be considerably reduced by using a low-resistivity epitaxial p-type substrate as the starting material, which can subsequently act as a very low resistance ground-plane to collect substrate currents.

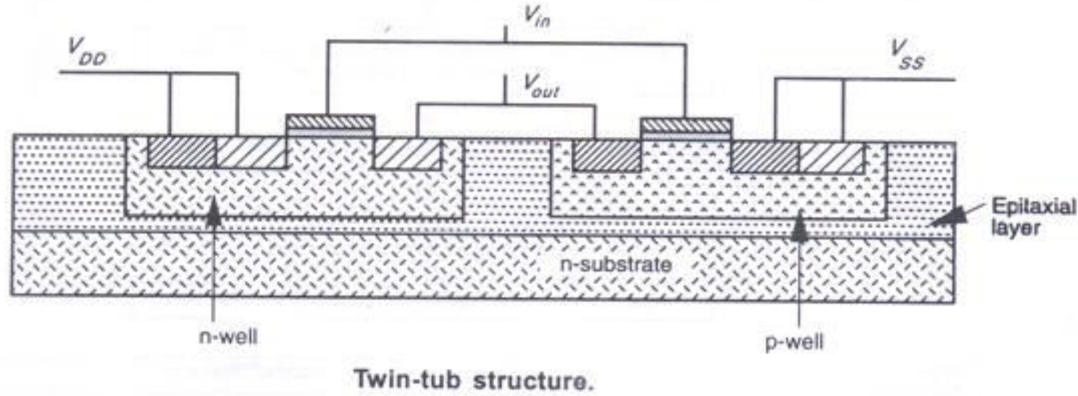
However, a factor of the n-well process is that the performance of the already poorly performing p-transistor is even further degraded. Modem process lines have come to grips with these problems, and good device performance may be achieved for both p-well and n-well fabrication.

**The Twin-Tub Process:**

A logical extension of the p-well and n-well approaches is the twin-tub fabrication process. Using twin tub technology, we can optimize NMOS and PMOS transistors separately. This means that transistor parameters such as threshold voltage, body effect and the channel transconductance of both types of transistors can be tuned independently.

A high resistivity n-type substrate, with a lightly doped epitaxial layer on top, forms the starting material for this technology. The n-well and p-well are formed on this epitaxial layer which forms the actual substrate. Through this process it is possible to preserve the performance of n-transistors without compromising the p-transistors. The dopant concentrations can be carefully optimized to produce the desired device characteristics because two independent doping steps are performed to create the well regions. This is particularly important as far as latch-up is concerned.

The conventional n-well CMOS process suffers from, among other effects, the problem of unbalanced drain parasitic since the doping density of the well region typically being about one order of magnitude higher than the substrate. This problem is absent in the twin-tub process. The below Figure will shows an inverter circuit fabricated by the Twin well process.



**Twin-tub structure.**

**Silicon on Insulator (SOI)**

To improve process characteristics such as speed and latch-up susceptibility, technologists have sought to use an insulating substrate instead of silicon as the substrate material. Completely isolated NMOS and PMOS transistors can be created virtually side by side on an insulating substrate (eg. sapphire) by using the SOI CMOS technology.

This technology offers advantages in the form of higher integration density (because of the absence of well regions), complete avoidance of the latch-up problem, and lower parasitic capacitances compared to the conventional n-well or twin-tub CMOS processes. But this technology comes with the disadvantage of higher cost than the standard n-well CMOS process. Yet the improvements of device performance and the absence of latch-up problems can justify its use, especially in deep submicron devices.

**BI-CMOS TECHNOLOGY:**

A BiCMOS circuit consists of both bipolar junction transistors and MOS transistors on a single substrate. The deficiency of MOS technology is the limited load driving capabilities of MOS transistors. This is due to the limited current sourcing and current sinking abilities associated with both p- and n-transistors and although it is possible, to design so-called super buffers using MOS transistors alone, such arrangements do not always compare well with the capabilities of bipolar transistors. Bipolar transistors also provide higher gain and have better noise and high frequency characteristics than MOS transistors.

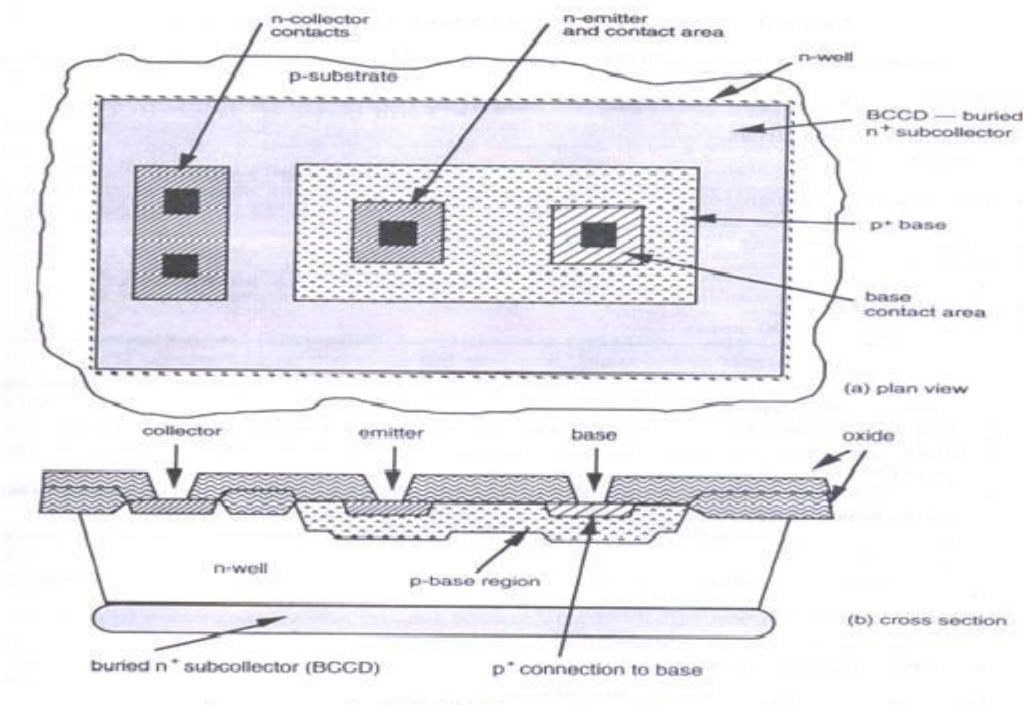
To drive large capacitive loads Bi-CMOS technology is used. As this technology combines Bipolar and CMOS transistors in a single integrated circuit, it has the advantages of both bipolar and CMOS transistors. BiCMOS is able to achieve VLSI circuits with speed-power-density performance previously not possible with either technology individually.

Using BiCMOS gates may be an effective way of speeding up VLSI circuits. However, the application of BiCMOS in subsystems such as ALU, ROM, a register-file, a barrel shifter is not always an effective way of improving speed. This is because most gates in such structures do not have to drive large capacitive loads so that the BiCMOS arrangements give no speed advantage. To take advantage of BiCMOS, the whole functional entity, not just the logic gates, must be considered. A comparison between the characteristics of CMOS and bipolar circuits is set out in Table shown below.

**Comparison between CMOS and Bipolar Technologies**

| MOS Technology | Bipolar Technology |
|---|---|
| • Low static power dissipation | • High power dissipation |
| • High input impedance (low drive current) | • Low input impedance (high drive current) |
| • Scalable threshold voltage | ----- |
| • High nose margin | • Low voltage swing logic |
| • High packing density | • Low packing density |
| • High delay sensitivity to load ( fan-out limitations) | • Low delay sensitivity to load |
| • Low output drive current | • High output drive current |
| • Low $g_m$ | • High $g_m$ |
| | • High $f_T$ at low currents |
| • Bidirectional capability (drain and source are interchangeable) | • Essentially unidirectional |
| • A near ideal switching device | |

Theoretically there should be little difficulty in extending CMOS fabrication processes to include bipolar as well as MOS transistors. In fact, a problem of p-well and n-well CMOS processing is that parasitic bipolar transistors are inadvertently formed as part of the outcome of fabrication. The production of npn bipolar transistors with good performance characteristics can be achieved, for example, by extending the standard n-well CMOS processing to include further masks to add two additional layers- the n+ subcollector and p+ base layers. The npn transistor is formed in an n-well and the additional $p^+$ base region is located in the well to form the p-base region of the transistor. The second additional layer, the buried $n^+$ subcollector (BCCD), is added to reduce the n-well (collector) resistance and thus improve the quality of the bipolar transistor. The simplified general arrangement of such a bipolar npn transistor is shown in below figure.



Arrangement of BiCMOS npn transistor

**BiCMOS Fabrication in an N-well Process**

The basic process steps used are those already outlined for CMOS but with additional process steps and additional masks defining (i) the $p^+$ base region; (ii) $n^+$ collector area; and (iii) the buried subcollector (BCCD).

Below Table sets out the process steps for a single poly, single metal CMOS n-well process, showing the additional process steps for the bipolar devices.

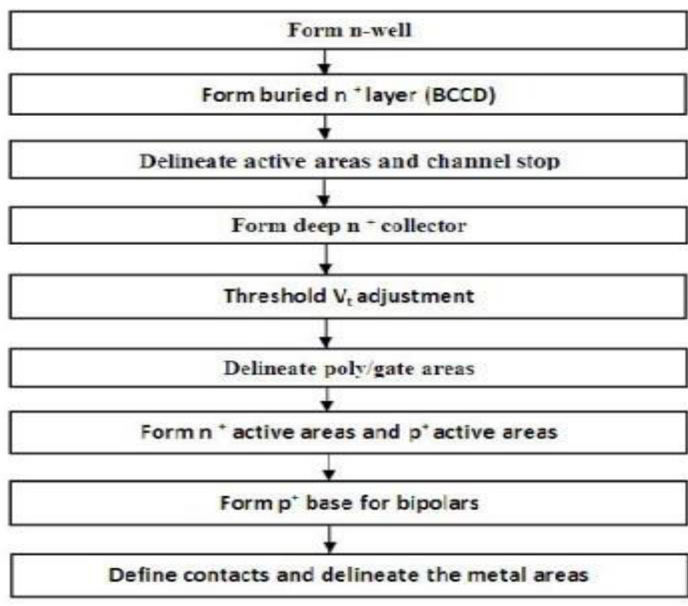
## N-Well BiCMOS fabrication Process Steps

Single poly, Single Metal CMOS

- Form n-well
- Define active area
- Channel stop
- Threshold Vt adjustment
- Define poly gate areas
- Form n+ active area
- Form p+ active area
- Define contacts
- Define the metal areas

Additional steps for Bipolar Devices

- Form buried n+ layer(BCCD)
- Form deep n+ collector
- Form p+ base for bipolar




n-well BiCMOS fabrication steps

### Some Aspects of Bipolar and CMOS Devices

There are several advantages if the properties of CMOS and bipolar technologies could be combined. This is achieved to a significant extent in the BiCMOS technology. As in all things, there is a penalty which, arises from the additional process steps, some loss of packing density and thus higher cost.

A further advantage which arises from BiCMOS technology is that analog amplifier design is facilitated and improved. High impedance CMOS transistors may be used for the input circuitry while the remaining stages and output drivers are realized using bipolar transistors.

To take maximum advantage of available silicon technologies one might envisage the following mix of technologies in a silicon system:

| | |
|---|---|
| CMOS | for logic |
| BiCMOS | for I/O and driver circuits |
| ECL | for critical high speed parts of the system |

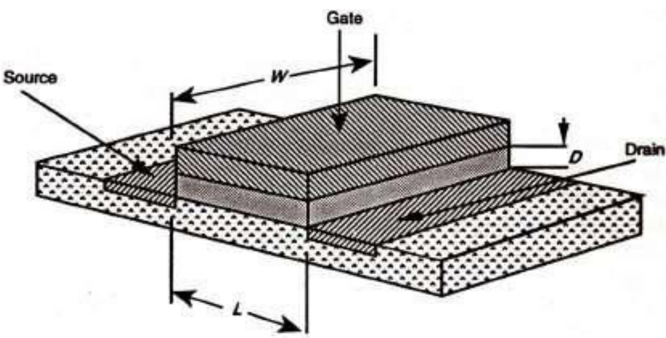
**BASIC ELECTRICAL PROPERTIES OF MOS ANA BiCMOS CIRCUITS**

**Drain-to-Source Current $I_{ds}$ Versus Voltage $V_{ds}$ Relationships:**

The working of a MOS transistor is based on the principle that the use of a voltage on the gate to induce a charge in the channel between source and drain, which may then be caused to move from source to drain under the influence of an electric field created by voltage Vds applied between drain and source. Since the charge induced is dependent on the gate to source voltage Vgs then Ids is dependent on both Vgs and Vds. Let us consider the diagram below in which electrons will flow source to drain .So, the drain current is given by

$$Ids = -Isd = \frac{\text{Charge induced in channel (Qc)}}{\text{Electron transit time}(\tau)}$$

Where the transit time is given by $\tau_{sd} = \frac{\text{Length of the channel}}{\text{Velocity (v)}}$




Where

$\mu$ =electron or hole mobility and $E_{ds}$ = Electric field

also , $E_{ds} = Vds/L$

**so,** $v = \mu.V_{ds}/L$

and $\tau_{ds} = L^2 / \mu.V_{ds}$

The typical values of μ at room temperature are given below.

$\mu_n \doteq 650\ cm^2/V\ sec$ (surface)

$\mu_p \doteq 240\ cm^2/V\ sec$ (surface)

**The Non-saturated Region :**

Let us consider the Id vs Vd relationships in the non-saturated region .The charge induced in the channel due to gate voltage is duo to the voltage difference between the gate and the channel, Vgs (assuming substrate connected to source). The voltage along the channel varies linearly with distance *X* from the source due to the IR drop in the channel. In the non-saturated state the average value is Vds/2. Also the effective gate voltage Vg = Vgs – Vt where Vt, is the threshold voltage needed to invert the charge under the gate and establish the channel.

Note: the charge/unit area = Eg $\varepsilon_{ins}\varepsilon o$.

Hence the induced charge is Qc = Eg $\varepsilon_{ins}$ εoW. L

Where $E_G$= average electric field gate to channel

$\varepsilon_{ins}$= relative permittivity of insulation between gate and channel (≈ 4.0 for silicon dioxide)

$\varepsilon_0$ = permittivity of free space ($8.85 * 10^{-14}\ Fcm^{-1}$)

$E_g = [(V_{gs}- V_t) – V_{ds}/2] / D$

where D= oxide thickness

**Threshold Voltage**

The voltage at which the surface of the semiconductor gets inverted to the opposite polarity is known as threshold voltage. At the threshold voltage condition, the concentration of electrons / holes accumulated near the surface in an n MOS / p MOS is equal to the doping concentration of the bulk doping concentration.

$V_t$ for n MOS→ +vei.e$V_{gs}$>$V_{tn}$

$V_t$ for p MOS→ -vei.e$V_{gs}$<$V_{tp}$

The threshold voltage of a MOSFET is defined as the value of the gate to source voltage which is sufficient to produce a surface inversion layer when $V_{DS}$= 0.

(or)

The voltage at which the surface of the semiconductor gets inverted to opposite polarity is known as **Threshold Voltage** ($V_t$).

$$Q_C = \frac{WL\,\varepsilon_{ins}\varepsilon_0}{D}[(V_{gs} - V_t) - V_{ds}/2]$$

So, by combining the above two equations ,we get

$$I_{ds} = Q_C / \tau ds$$

$$= \frac{WL\,\varepsilon_{ins}\varepsilon_0}{D}[V_{gs} - V_t) - V_{ds}/2] \Big/ \frac{L^2}{\mu V_{DS}}$$

$$I_{ds} = \frac{\varepsilon_{ins}\varepsilon_0\mu}{D}\frac{W}{L}[V_{gs} - V_t) - V_{ds}^2/2]$$

$$I_{ds} = K \frac{W}{L} [V_{gs} - V_t) - V_{ds}^2/2]$$

In the non-saturated or resistive region where

$$V_{ds} < V_{gs} - V_t \text{ and}$$

$$K = \frac{\varepsilon_{ins}\varepsilon_0\mu}{D}$$

The factor W/L is geometric factor $\beta = K\, W/L$

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

Gate / channel capacitance $C_g = \frac{WL\,\varepsilon_{ins}\varepsilon_0}{D}$

$$K = \frac{C_g\mu}{WL}$$

$$I_{ds} = \frac{C_g\,\mu}{L^2} [V_{gs} - V_t) - V_{ds}^2/2]$$

$$C_g = C_0 WL$$

$$I_{ds} = C_0\,\mu\frac{W}{L} [V_{gs} - V_t) - V_{ds}^2/2]$$

**The Saturated Region:**

Saturation begins when $V_{ds} = V_{gs} - V_t$, the IRdrop in the channel equals the effective gate to channel voltage at the drain and assume that the current remains fairly constant as $V_{ds}$ increases further.

$$I_{ds} = K \frac{W}{L} (V_{gs} - V_t)^2/2$$

$$I_{ds} = \beta/2\,(V_{gs} - V_t)^2$$
$$I_{ds} = \frac{C_g\mu}{2L^2}(V_{gs} - V_t)^2$$
$$I_{ds} = C_0\mu\,\frac{W}{2L}\,(V_{gs} - V_t)^2$$

$I_{ds}$ for both enhancement and depletion mode devices, the threshold voltage for the n MOS depletion mode device (denoted $V_{td}$is -ve).

**Aspects of MOS Transistor Threshold Voltage Vt :**

The gate structure of a MOS transistor consists, of charges stored in the dielectric layers and in the surface to surface interfaces as well as in the substrate itself. Switching an enhancement mode MOS transistor from the off to the on state consists in applying sufficient gate voltage to neutralize these charges and enable the underlying silicon to undergo an inversion due to the electric field from the gate. Switching a depletion mode nMOS transistor from the on to the off state consists in applying enough voltage to the gate to add to the stored charge and invert the 'n' implant region to 'p'. The threshold voltage Vt may be expressed as:

$$V_t = \phi_{ms} \; \frac{Q_B - Q_{SS}}{C_0} + 2\phi_{fN}$$

where $Q_B$ = the charge per unit area in the depletion layer below the oxide
Qss = charge density at Si: SiO2 interface
Co =Capacitance per unit area.
Φms = work function difference between gate and Si
$\Phi_{fN}$ = Fermi level potential between inverted surface and bulk Si
For polynomial gate and silicon substrate, the value of Φms is negative but negligible and the magnitude and sign of Vt are thus determined by balancing the other terms in the equation.

To evaluate the Vt the other terms are determined as below.

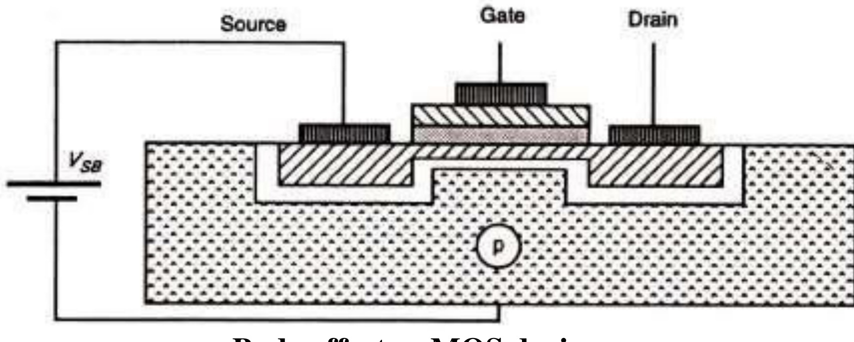
$$Q_B = \sqrt{2\varepsilon_0\varepsilon_{Si} q N (2\phi_{fN} + V_{SB})}\ \text{coulomb/m}^2$$

$$\phi_{fN} = \frac{kT}{q}\ \ln\frac{N}{n_i}\ \text{volts}$$

$$Q_{SS} = (1.5\ \text{to}\ 8) \times 10^{-8}\ \text{coulomb/m}^2$$

**Body Effect :**

Generally while studying the MOS transistors it is treated as a three terminal device. But ,the body of the transistor is also an implicit terminal which helps to understand the characteristics of the transistor. Considering the body of the MOS transistor as a terminal is known as the body effect. The potential difference between the source and the body (Vsb) affects the threshold voltage of the transistor. In many situations, this Body Effect is relatively insignificant, so we can (unless **otherwise** stated) ignore the Body Effect. But it is not always insignificant, in some cases it can have a tremendous impact on MOSFET circuit performance.




**Body effect - nMOS device**

Increasing Vsb causes the channel to be depleted of charge carriers and thus the threshold voltage is raised. Change in Vt is given by ΔVt = γ.(Vsb)1/2 where γ is a constant which depends on substrate doping so that the more lightly doped the substrate, the smaller will be the body effect

The threshold voltage can be written as

$$V_t = V_t(0) + \left(\frac{D}{\varepsilon_{ins}\varepsilon_0}\right)\sqrt{2\varepsilon_0\varepsilon_{Si}QN}\,.\,(V_{SB})^{1/2}$$

Where Vt(0) is the threshold voltage for Vsd = 0
For n-MOS depletion mode transistors ,the body voltage values at different VDD voltages are given below.

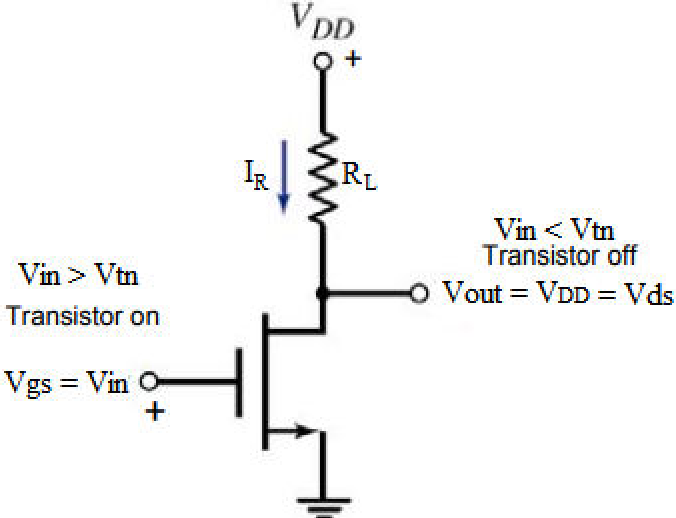
$V_{SB}$ = 0 V ; Vsd = -0.7$V_{DD}$ (= - 3.5 V for $V_{DD}$ =+5V )
$V_{SB}$ = 5 V ; Vsd = -0.6$V_{DD}$ (= - 3.0 V for $V_{DD}$ =+5V )

**The nMOS INVERTER :** For any IC technology used in digital circuit design, the basic circuit element is the logic inverter. Once the operation and characterization of an inverter circuits are thoroughly understood, the results can be extended to the design of the logic gates and other more complex circuits.

An inverter circuit is a very important circuit for producing a complete range of logic circuits. This is needed for restoring logic levels, for Nand and Nor gates, and for sequential and memory circuits of various forms.

**nMOS INVERTER with Resistive Load:**

A simple inverter circuit can be constructed using a transistor with source connected to ground and a load resistor of connected from the drain to the positive supply rail $V_{DD}$. The output is taken from the drain and the input applied between gate and ground. The basic structure of a resistive load inverter is shown in the figure given below.




**Circuit Operation :** Here, enhancement type nMOS acts as the driver transistor. The load consists of a simple linear resistor $R_L$. When the input of the driver transistor is less than threshold voltage $V_{tn}$ ($V_{in} < V_{tn}$), driver transistor is in the cut – off region and does not conduct any current. So, the voltage drop across the load resistor is ZERO and output voltage is equal to the $V_{DD}$.

Now, when the input voltage increases slightly above Vtn, driver transistor will start conducting the non-zero current and goes in saturation region since Vds > (Vgs – Vtn) **.**

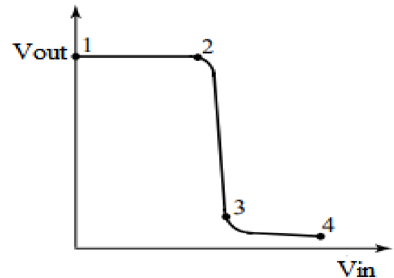
$$V_{out} = V_{DD} - i_R R_L$$

$$I_R = I_{ds} = [\beta(V_{gs} - V_{tn})^2]/2$$

Increasing the input voltage further, driver transistor will enter into the linear region since Vds < (Vgs – Vtn) and output of the driver transistor decreases.

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

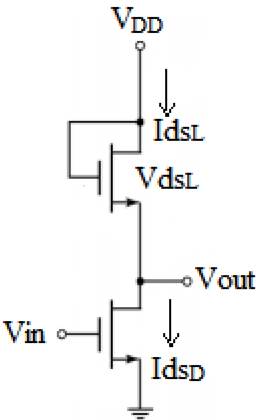
VTC of the resistive load inverter, shown below.



But, during the fabrication resistors are not conveniently produced on the silicon substrate and even small values of resistors occupy excessively large areas .Hence some other form of load resistance is used. A more convenient way to solve this problem is to use a MOS transistor as the load.

**Enhancement Load NMOS:**

An n-channel enhancement-mode MOSFET with the gate connected to the drain can be used as load device in an NMOS inverter. Since the gate and drain of the transistor are connected, we have Vgs = Vds. When Vgs=Vds >Vtn, a non zero drain current is induced in the transistor and thus the transistor operates in saturation only. And following condition is satisfied Vds > (Vds - Vt). The inverter with enhancement-type load device is shown in the figure.




When Vin < $Vtn_D$ , the driver is cut off and the drain currents are zero. It means

$Ids_L = 0 = [\beta_L(Vds_L - Vtn_L)^2]/2$

So $Vds_L – Vtn_L = 0$

But $Vds_L = V_{DD} –Vout$

$V_{DD} –Vout – Vtn_L = 0$

$Vout = V_{DD} –Vtn_L$

When Vin > $Vtn_D$ , the driver turns on and is biased in saturation region.

So $Ids_L = Ids_D$

$[\beta_D(Vgs_D – Vtn_D)^2]/2 = [\beta_L(Vds_L - Vtn_L)^2]/2$

$[\beta_D(Vin – Vtn_D)^2]/2 = [\beta_L(V_{DD} –Vout - Vtn_L)^2]/2$

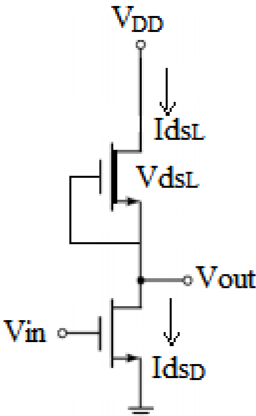
$Vout = V_{DD} - Vtn_L - \{\sqrt{\beta_L} /\beta_D\}(Vin – Vtn_D)$

As the Vin increases , the Vout decreases linearly with vin.

Increasing the input voltage further, driver transistor will enter into the linear region since $Vds_D < (Vgs_D – Vtn_D)$ and output of the driver transistor decreases.

The main drawback of this inverter is Vout limited to $V_{DD} –Vtn_L$.

**NMOS Inverter with Depletion Load:** This is an alternate form of the NMOS inverter that uses an **depletion-mode MOSFET load** device with gate and source terminal connected. This inverter has the advantage of Vout = $V_{DD}$.




The salient features of the n-MOS depletion mode transistor are:

- In n- channel depletion mode MOSFET, an n-channel region or inversion layer exists under the gate oxide layer even at zero gate voltage and hence term depletion mode**.**
- A negative voltage must be applied to the gate to turn the device off**.**
- The threshold voltage is always negative for this kind of device.

The salient features of the n-MOS inverter are

- For the depletion mode transistor, the gate is connected to the source(Vgs = 0) so it is always on .
- In this configuration the depletion mode device is called the pull-up (P.U) and the enhancement mode device the pull-down (P.D) transistor.
  - With no current drawn from the output, the currents Ids for both transistors must be equal.

When Vin < $Vtn_D$ , the driver is cut off and no drain current conduct in either transistor. That means the load transistor must be in the linear region of the operation and the output current can be expressed as fellows

$Ids_L = 0 = \beta_L[(Vgs_L - Vtn_L)Vds_L - \{ Vds_L^2/2\}]$

But $Vgs_L = 0$

$Ids_L = 0 = - \beta_L\, Vds_L\, [\, Vtn_{L+} \{Vds_L/2\}]$ which gives $Vds_L = 0$

But $Vds_L = V_{DD} - Vout$

$V_{DD} - Vout = 0$

$Vout = V_{DD}$

When Vin > $Vtn_D$ , the driver turns on and is biased in saturation region. However load is in non saturation region.so that

$Ids_L = \beta_L[(Vgs_L - Vtn_L)Vds_L - \{ Vds_L^2/2\}]$

$= \beta_L[(0 - Vtn_L) (V_{DD} - Vout) - \{ (V_{DD} - Vout)^2/2\}]$

$Ids_D = [\beta_D(Vgs_D - Vtn_D)^2]/2$

$= [\beta_D(Vin - Vtn_D)^2]/2$

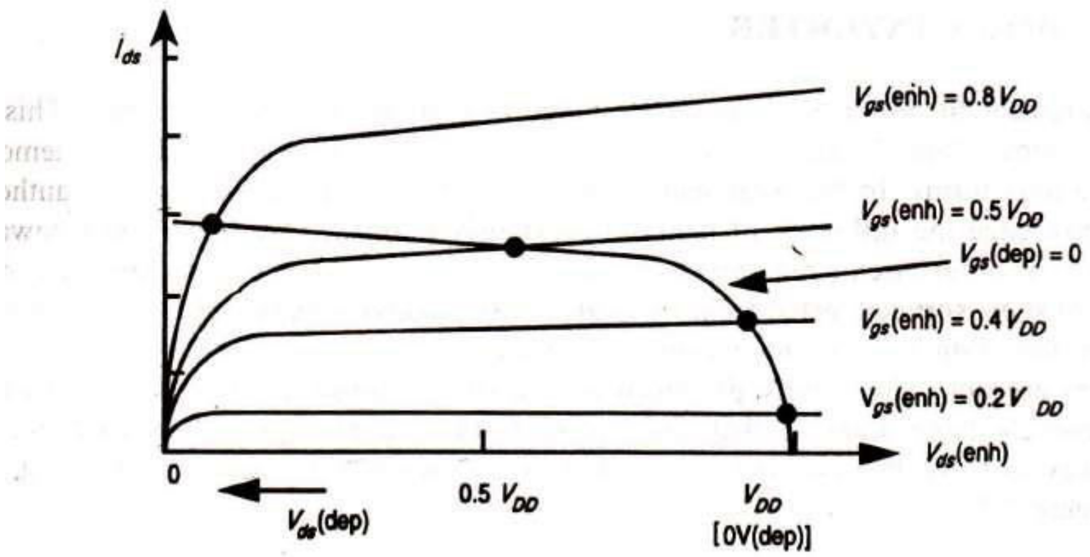
By euqting above two equations we have a non linear relation between Vout and Vin.

Increasing the input voltage further, both the transistors will enter into the saturation region. Then the relation between Vout and Vin is linear.

As Increasing the input voltage further and further, driver transistor biased in the non-saturation region while the load is in the saturation. This implies that input and output voltages are not linear in this region.

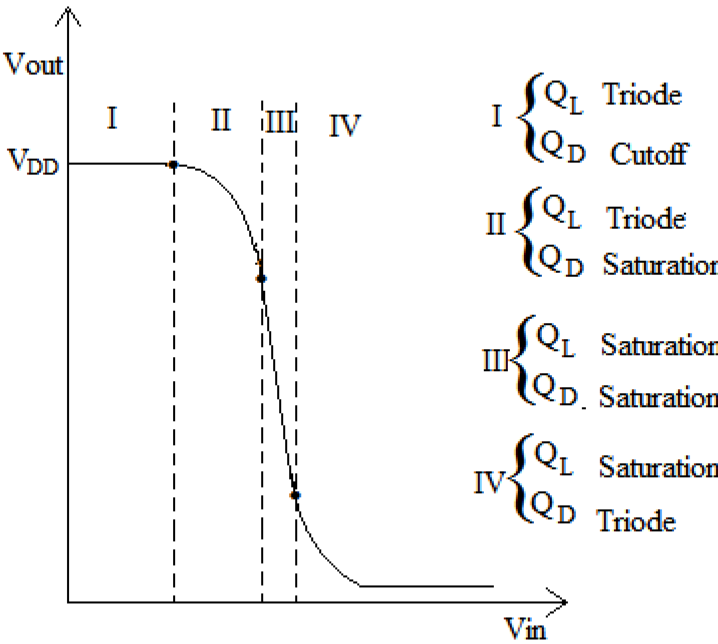
**nMOS Inverter transfer characteristic**.

The transfer characteristic is drawn by taking Vds on x-axis and Ids on Y-axis for both enhancement and depletion mode transistors. So,to obtain the inverter transfer characteristic for *Vgs* = 0 depletion mode characteristic curve is superimposed on the family of curves for the enhancement mode device and from the graph it can be seen that , maximum voltage across the enhancement mode device corresponds to minimum voltage across the depletion mode transistor.




From the graph it is clear that as Vin(=Vgs p.d. transistor) exceeds the Pull down threshold voltage current begins to flow. The output voltage Vout thus decreases and the subsequent increases in $V_{in}$ will cause the Pull down transistor to come out of saturation and become resistive.

**Inverter voltage transfer characteristic:**




**Determination of Pull-up to Pull –Down Ratio** ***(Zp.u}Zp.d.)*****for an nMOS Inverter driven by another nMOS Inverter :**

Let us consider the arrangement shown in Fig.(a). in which an inverter is driven from the output of another similar inverter. Consider the depletion mode transistor for which Vgs = 0 under all conditions, and also assume that in order to cascade inverters without degradation the condition

$$V_{in} = V_{out} = V_{inv}$$

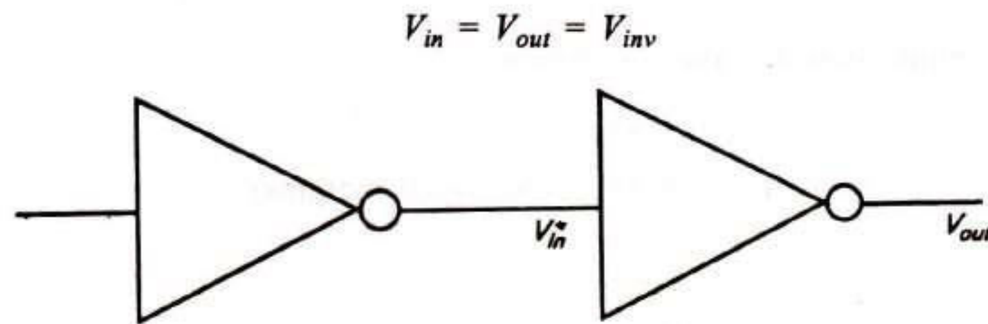



Fig.(a).Inverter driven by another inverter.

For equal margins around the inverter threshold, we set Vinv = 0.5VDD · At this point both transistors are in saturation and we can write that

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

In the depletion mode $I_{ds} = K\frac{W_{p.u.}}{L_{p.u.}}\frac{(-V_{td})^2}{2}$ since $V_{gs} = 0$

and in the enhancement mode

$$I_{ds} = K\frac{W_{p.d.}}{L_{p.d.}}\frac{(V_{inv} - V_t)^2}{2} \text{ since } V_{gs} = V_{inv}$$

Equating (since currents are the same) we have

$$\frac{W_{p.d.}}{L_{p.d.}}(V_{inv} - V_t)^2 = \frac{W_{p.u.}}{L_{p.u.}}(-V_{td})^2$$

where Wp.d , Lp.d , Wp.u. and Lp.u are the widths and lengths of the pull-down and pull-up transistors respectively.

So,we can write that

$$Z_{p.d.} = \frac{L_{p.d.}}{W_{p.d.}};\ Z_{p.u.} = \frac{L_{p.u.}}{W_{p.u.}}$$

we have

$$\frac{1}{Z_{p.d.}}(V_{inv} - V_t)^2 = \frac{1}{Z_{p.u.}}(-V_{td})^2$$

whence

$$V_{inv} = V_t - \frac{V_{td}}{\sqrt{Z_{p.u.}/Z_{p.d.}}}$$

The typical, values for Vt ,Vinv and Vtd are

$$V_t = 0.2V_{DD};\ V_{td} = -0.6V_{DD}$$

$$V_{inv} = 0.5V_{DD} \text{ (for equal margins)}$$

Substituting these values in the above equation ,we get

$$0.5 = 0.2 + \frac{0.6}{\sqrt{Z_{p.u.}/Z_{p.d.}}}$$

Here

$$\sqrt{Z_{p.u.}/Z_{p.d.}} = 2$$

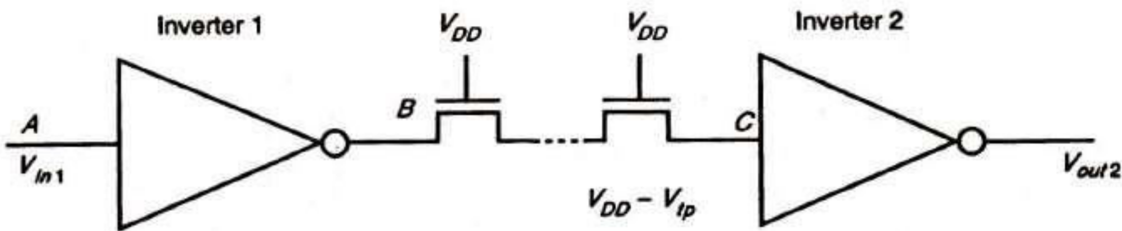
So,we get

$$Z_{p.u.}/Z_{p.d.} = 4/1$$

This is the ratio for pull-up to pull down ratio for an inverter directly driven by another inverter.

**Pull -Up to Pull-Down ratio for an nMOS Inverter driven through one or more Pass Transistors**

Let us consider an arrangement in which the input to inverter 2 comes from the output of inverter 1 but passes through one or more nMOS transistors as shown in Fig. below (These transistors are called pass transistors).




The connection of pass transistors in series will degrade the logic 1 level / into inverter 2 so that the output will not be a proper logic 0 level. The critical condition is , when point A is at 0 volts and B is thus at VDD. but the voltage into inverter 2at point C is now reduced from VDD by the threshold voltage of the series pass transistor. With all pass transistor gates connected to VDD there is a loss of Vtp, however many are connected in series, since no static current flows through them and there can be no voltage drop in the channels. Therefore, the input voltage to inverter 2 is

$$V_{in2} = V_{DD} - V_{tp}$$

where Vtp = threshold voltage for a pass transistor.

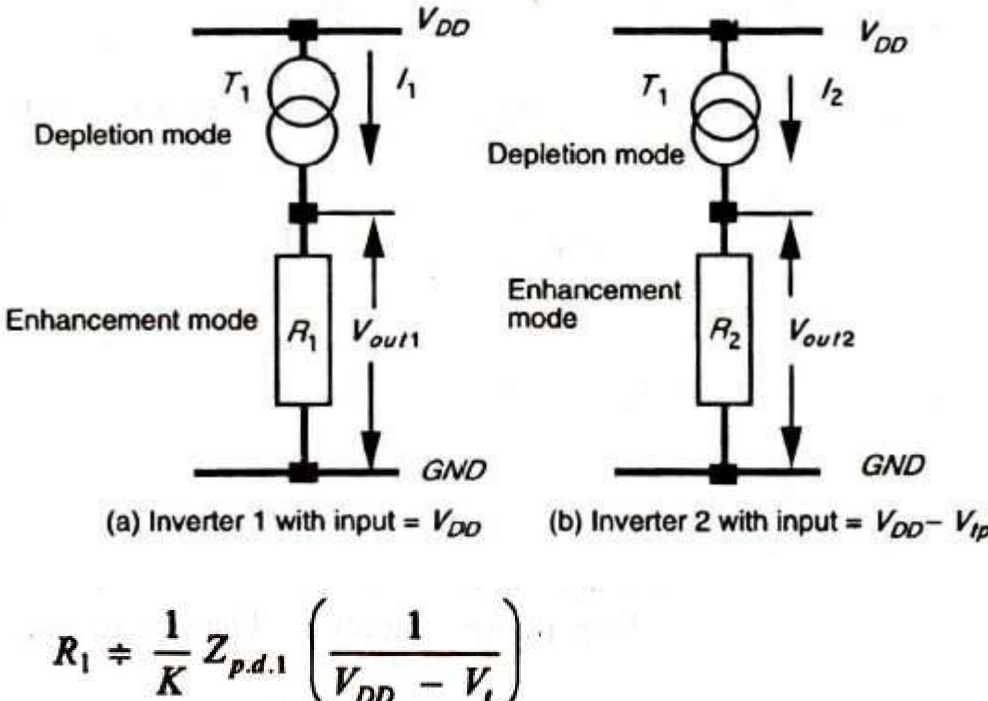
Let us consider the inverter 1 shown in Fig.(a) with input = $V_{DD}$. If the input is at $V_{DD}$, then the pull-down transistor T2 is conducting but with a low voltage across it; therefore, it is in its resistive region represented by R1 in Fig.(a) below. Meanwhile, the pull up transistor T1 is in saturation and is represented as a current source.

For the pull down transistor

$$R_1 = \frac{V_{ds1}}{I_{ds}} = \frac{1}{K}\frac{L_{p.d.1}}{W_{p.d.1}}\left(\frac{1}{V_{DD} - V_t - \frac{V_{ds1}}{2}}\right)$$

$$I_{ds} = K\frac{W_{p.d.1}}{L_{p.d.1}}\left((V_{DD} - V_t)\,V_{ds1} - \frac{V_{ds1}^2}{2}\right)$$

Since Vds is small, Vds/2 can be neglected in the above expression.



(a) Inverter 1 with input = $V_{DD}$ (b) Inverter 2 with input = $V_{DD} - V_{tp}$

$$R_1 \doteq \frac{1}{K} Z_{p.d.1}\left(\frac{1}{V_{DD} - V_t}\right)$$

Now, for depletion mode pull-up transistor in saturation with Vgs = 0

$$I_1 = I_{ds} = K\frac{W_{p.u.1}}{L_{p.u.1}}\,\frac{(-V_{td})^2}{2}$$

The product $I_1R_1 = V_{out1}$

$$V_{out1} = I_1R_1 = \frac{Z_{p.d.1}}{Z_{p.u.1}}\left(\frac{1}{V_{DD} - V_t}\right)\frac{(V_{td})^2}{2}$$

Let us now consider the inverter 2 Fig.b .when input = $V_{DD}$- Vtp.

$$R_2 \doteqdot \frac{1}{K} Z_{p.d.2} \frac{1}{((V_{DD} - V_{tp}) - V_t}$$

$$I_2 = K \frac{1}{Z_{p.u.2}} \frac{(-V_{td})^2}{2}$$

Whence,

$$V_{out\,2} = I_2 R_2 = \frac{Z_{p.d.2}}{Z_{p.u.2}} \left( \frac{1}{V_{DD} - V_{tp} - V_t} \right) \frac{(-V_{td})^2}{2}$$

If inverter 2 is to have the same output voltage under these conditions then $V_{out1} = V_{out2}$. That is

$I_1R_1=I_2R_2$ , therefore

$$\frac{Z_{p.u.2}}{Z_{p.d.2}} = \frac{Z_{p.u.1}}{Z_{p.d.1}} \frac{(V_{DD} - V_t)}{(V_{DD} - V_{tp} - V_t)}$$

Considering the typical values

$$V_t = 0.2 V_{DD}$$

$$V_{tp} = 0.3 V_{DD}^*$$

$$\frac{Z_{p.u.2}}{Z_{p.d.2}} = \frac{Z_{p.u.1}}{Z_{p.d.1}} \frac{0.8}{0.2}$$

Therefore

$$\frac{Z_{p.u.2}}{Z_{p.d.2}} \doteqdot 2 \frac{Z_{p.u.1}}{Z_{p.d.1}} = \frac{8}{1}$$

From the above theory it is clear that, for an n-MOS transistor

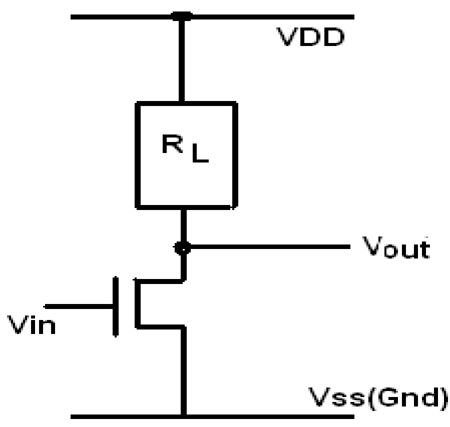
(i). An inverter driven directly from the output of another should have a $Z_{p.u}/ Z_{pd.}$ ratio of $\geq 4/1$.

(ii).An inverter driven through one or more pass transistors should have a Zp.u./Zp.d ratio of ≥8/1

**ALTERMTIVE FORMS OF PULL –UP**

Generally the inverter circuit will have a depletion mode pull-up transistor as its load. But there are also other configurations .Let us consider four such arrangements.

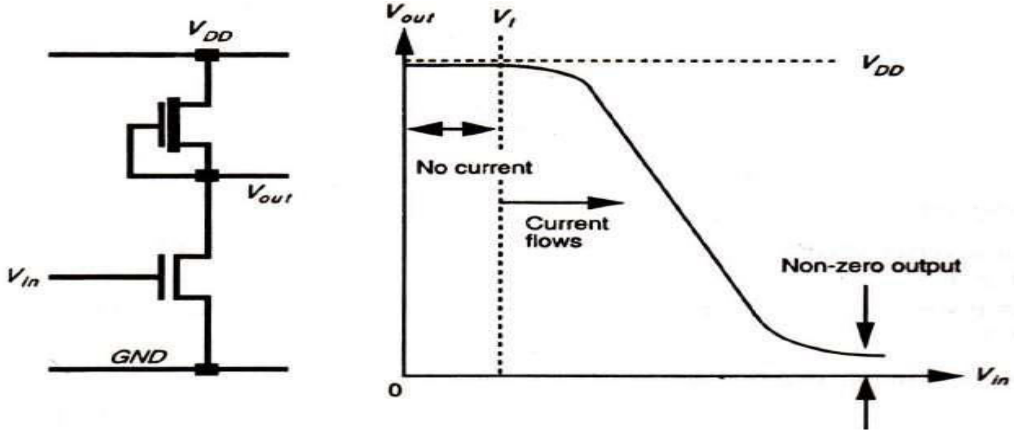
**(i).Load resistance RL :** This arrangement consists of a load resistor as a pull-up as shown in the diagram below. But it is not widely used because of the large space requirements of resistors produced in a silicon substrate.




**2. nMOS depletion mode transistor pull-up :** This arrangement consists of a depletion mode transistor as pull-up. The arrangement and the transfer characteristic are shown below.In this type of arrangement we observe

(a) Dissipation is high , since rail to rail current flows when Vin = logical 1.

(b) Switching of output from 1 to 0 begins when Vin exceeds Vt, of pull-down device

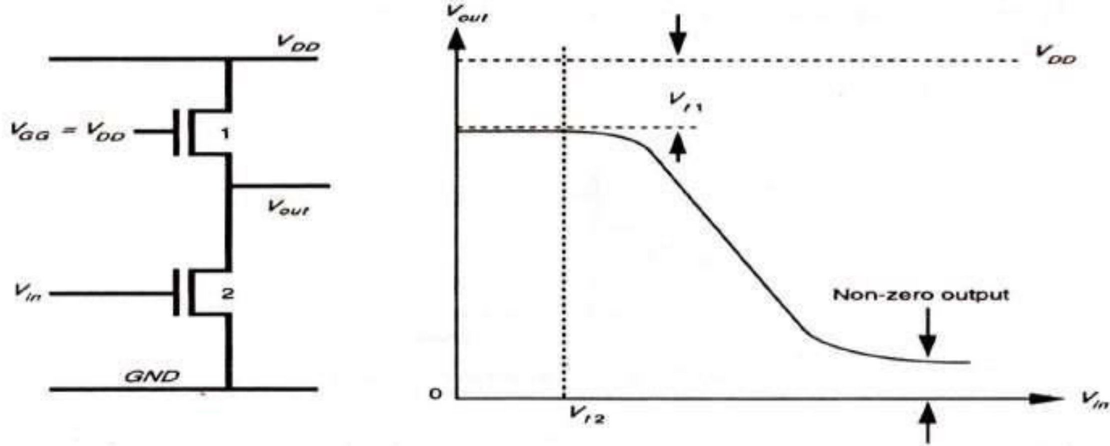



**nMOS depletion mode transistor pull-up and transfer characteristic**

(c) When switching the output from 1 to 0, the pull-up device is non-saturated initially and this

presents lower resistance through which to charge capacitive loads .

**3. nMOS enhancement mode pull-up :**This arrangement consists of a n-MOS enhancement mode transistor as pull-up. The arrangement and the transfer characteristic are shown below.




**nMOS enhancement mode pull-up and transfer characteristic**

The important features of this arrangement are

(a) Dissipation is high since current flows when Vin =logical 1 (VGG is returned to VDD) .

(b) Vout can never reach VDD (logical I) if VGG = VDD as is normally the case.

(c) VGG may be derived from a switching source, for example, one phase of a clock, so that dissipation can be greatly reduced.

(d) If VGG is higher than VDD then an extra supply rail is required.

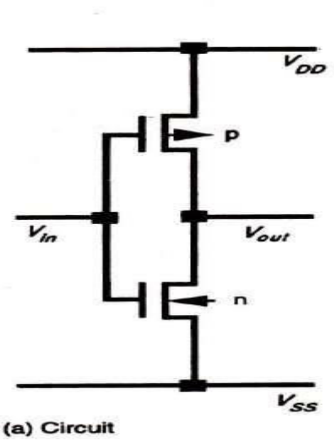
**4. Complementary transistor pull-up (CMOS) :** This arrangement consists of a C-MOS arrangement as pull-up. The arrangement and the transfer characteristic are shown below

**The salient features of this arrangement are**

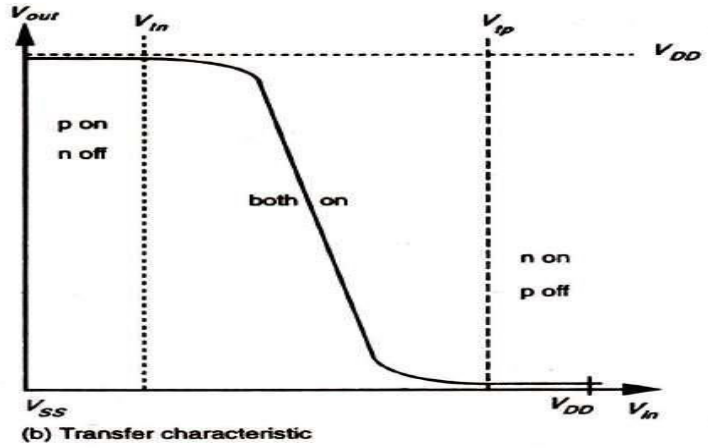
(a) No current flows either for logical 0 or for logical 1 inputs.

(b) Full logical 1 and 0 levels are presented at the output.

(c) For devices of similar dimensions the p-channel is slower than the n-channel device.



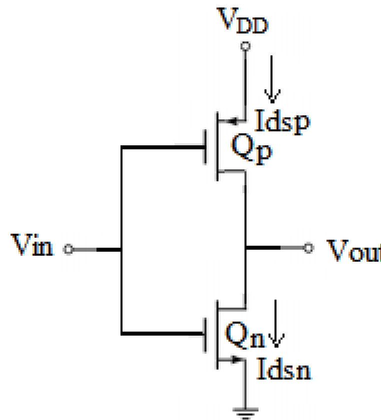

(a) Circuit




(b) Transfer characteristic

**CMOS Inverter :**

The inverter is the very important part of all digital designs. Once its operation and properties are clearly understood, Complex structures like NAND gates, adders, multipliers, and microprocessors can also be easily done. The electrical behavior of these complex circuits can be almost completely derived by extrapolating the results obtained for inverters. As shown in the diagram below the CMOS transistor is designed using p-MOS and n-MOS transistors




In the inverter circuit ,if the input is high .the lower n-MOS device closes to discharge the capacitive load .Similarly ,if the input is low,the top p-MOS device is turned on to charge the capacitive load .At no time both the devices are on ,which prevents the DC current flowing from positive power supply to ground. Qualitatively this circuit acts like the switching circuit, since the p-channel transistor has exactly the opposite characteristics of the n-channel transistor. In the transition region both transistors are saturated and the circuit operates with a large voltage gain.

Circuit operation:

The operation of CMOS inverter can be divided into five regions .The behavior of n- and p-devices in each of region is explained below.

**Region 1 :** This region is defined by 0 =< Vin < Vtn in which the n-device is cut off (Idsn =0), and the p-device is in the linear region. Since Idsn = –Idsp, the drain-tosource current Idsp for the p-device is also zero.

$Idsp = 0 = \beta p[(Vgsp - |Vtp|)Vdsp - \{ Vdsp^2/2\}]$

$= \beta p\ Vdsp\ [(Vgsp - |Vtp|) - \{ Vdsp /2\}]$

But $Vgsp = Vs - Vg = V_{DD} - Vin$ (By considering all positive voltages)

$Vdsp = Vs - Vd = V_{DD} - Vout$

Inorder to get Idsp = 0,Vdsp should be zero

$Vdsp = 0 = V_{DD} - Vout$

$Vout = V_{DD}$

**Region 2 :** This region is defined by $Vtn =< Vin < \{V_{DD}/2\}$ in which the n-device is biased in saturation ($Vgsn > Vtn$ and $Vdsn = Vout = V_{DD}$), while the p-device is in the linear region($Vdsp$ = small ).

$$Idsp = \beta p[(Vgsp - |Vtp|)Vdsp - \{ Vdsp^2/2\}]$$
$$Idsn = \beta n[(Vgsn - Vtn)^2]/2$$
$$Vgsn = Vin$$
$$Idsn = \beta n[(Vin - Vtn)^2/2$$

By euqting above two equations we have a non linear relation between Vout and Vin.

**Region 3 :** This region is defined by $Vin = V_{DD}/2$ in which Vin ) - |Vtp|]biased in saturation.

**To find the point at which pMOS enter into saturation:**

The transition point for pMOS at which it enter into saturation is given by

$$Vdsp = (Vgsp - |Vtp|) =[(V_{DD} - Vin) - |Vtp|]$$
$$\text{But } Vdsp = V_{DD} - Vout$$

By equating above two equations we get

$$Vout = Vin + |Vtp|$$

At this point of Vdsp pMOS transistor enter into saturation.

So Idsn = Idsp

$$\beta n[(Vin - Vtn)^2]/2 = -- [(VDD - Vin) - |Vtp|]^2/2$$

if $\beta n = \beta p$ and $Vtn = |Vtp|$

$$Vin - Vtn = V_{DD} - Vin - |Vtp|$$
$$2Vin = V_{DD}$$
$$Vin = V_{DD}/2$$

Therefore $Vin = V_{DD}/2$ is the point at which both the transistors enter into saturation.

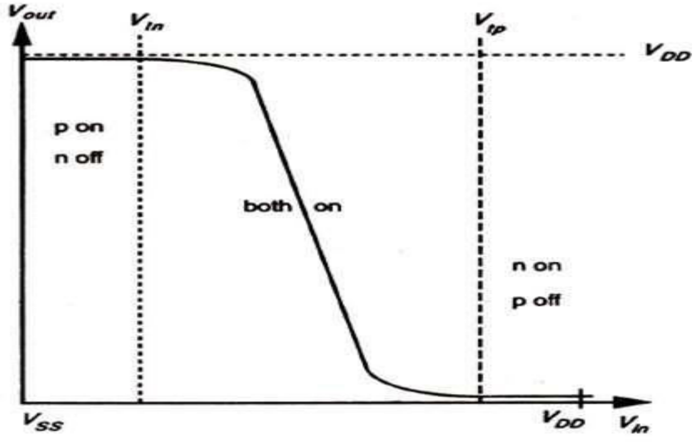
Which implies that **region 3** exists only for one value of Vin. We have assumed that a MOS device in saturation behaves like an ideal current soured with drain-to-source current being independent of Vds. In reality, as Vds increases, Ids also increases slightly; thus **region 3** has a finite slope. The significant factor to be noted is that in **region 3**, we have two current sources in series, which is an "unstable" condition. Thus a small input voltage as a large effect at the output. This makes the output transition very steep, which contrasts with the equivalent nMOS inverter characteristics. The above expression of Vin is particularly useful since it provides the basis for defining the gate threshold Vinv which corresponds to the state where Vout=Vin .This region also defines the "gain" of the CMOS inverter when used as a small signal amplifier.

**Region 4 :** This region is described by $V_{DD}/2 < Vin =< V_{DD} - |Vtp|$.

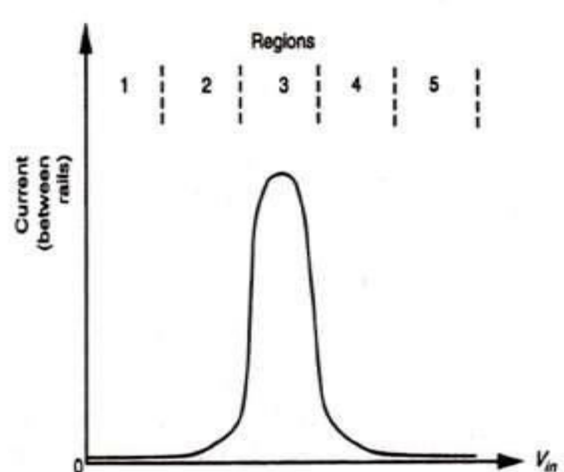
The p-device is in saturation while the n-device is operation in its nonsaturated region. In this region The relation between Vin and Vout is non linear.

**Region 5 :** This region is described by $Vin > V_{DD} - |Vtp|$ in which the p device is cut off ($Idsp = 0$), and the n-device is in the linear mode. Here, Vgsp= Vin - VDD Which is more positive than Vtp. The output in this region is Vout=0 From the transfer curve , it may be seen that the transition between the two states is very step. This characteristic is very desirable because the noise immunity is maximized. The gate-threshold voltage, Vinv, where Vin =Vout is dependent on $\beta n/\beta p$. Thus, for given proc kess, if we want to change $\beta n/\beta p$ we need to change the channel

dimensions, i.e.,channel-length L and channel-width W. Therefore it can be seen that as the ratio βn/βp is decreased, the transition region shifts from left to right; however, the output voltage transition remains sharp.
The CMOS transfer characteristic is shown in the below graph.




CMOS inverter Transfer characteristics

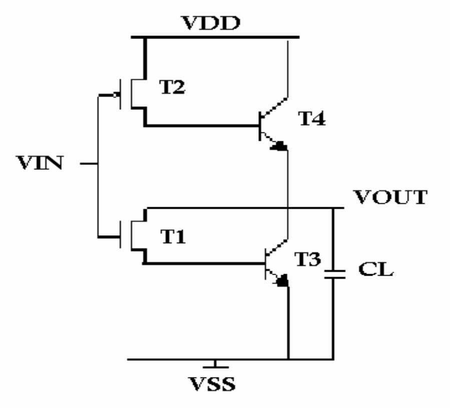



CMOS inverter current versus Vin plot

**The BiCMOS Inverter**

Two bipolar transistors (T3 and T4), one nMOS and one pMOS transistor (both enhancement-type devices, OFF at Vin=0V)
The MOS switches perform the logic function & bipolar transistors drive output loads
With Vin = 0, T1 is off therefore T3 is non-conducting T2 ON - supplies current to base of T4
T4 base voltage set to Vdd.T4 conducts & acts as current source to charge load CL towards Vdd.
Vout rises to Vdd - Vbe (of T4)/
With Vin = Vdd T2 is off therefore T4 is non-conducting.T1 is on and supplies current to the base of T3 thenT3 conducts & acts as a current sink to discharge load CL towards 0V.

Vout falls to 0V+ VCEsat (of T3)



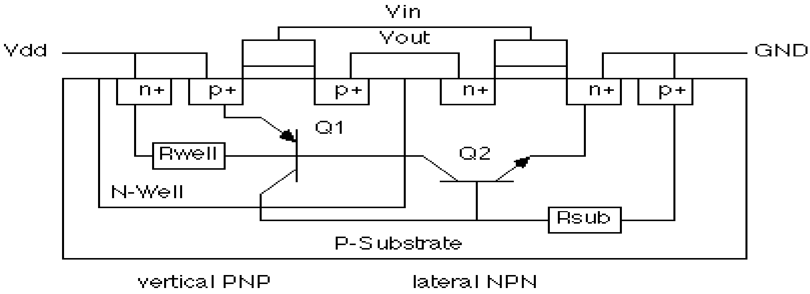

BiCMOS Inverter

- T3 & T4 present low impedances when turned on into saturation & load CL will be charged or discharged rapidly.
- Output logic levels will be good & will be close to rail voltages since VCEsat is quite small & VBE ≈ 0.7V. Therefore, inverter has high noise margins
- Inverter has high input impedance, i.e., MOS gate input
- Inverter has high drive capability but occupies a relatively small area

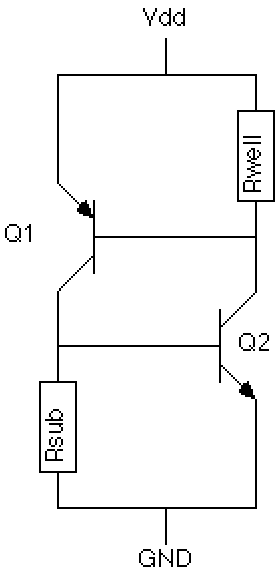
However, this is not a good arrangement to implement since no discharge path exists for current from the base of either bipolar transistor when it is being turned off, i.e. when Vin=Vdd, T2 is off and no conducting path to the base of T4 exists when Vin=0, T1 is off and no conducting path to the base of T3 exists

## Latch-up in CMOS circuits

A byproduct of the Bulk CMOS structure is a pair of parasitic bipolar transistors. The collector of each BJT is connected to the base of the other transistor in a positive feedback structure. A phenomenon called latch up can occur when (1) both BJT's conduct, creating a low resistance path between Vdd and GND **and** (2) the product of the gains of the two transistors in the feedback loop, b1 x b2, is greater than one. The result of latch up is at the minimum a circuit malfunction, and in the worst case, the destruction of the device.




Cross section of parasitic transistors in Bulk CMOS




Equivalent Circuit

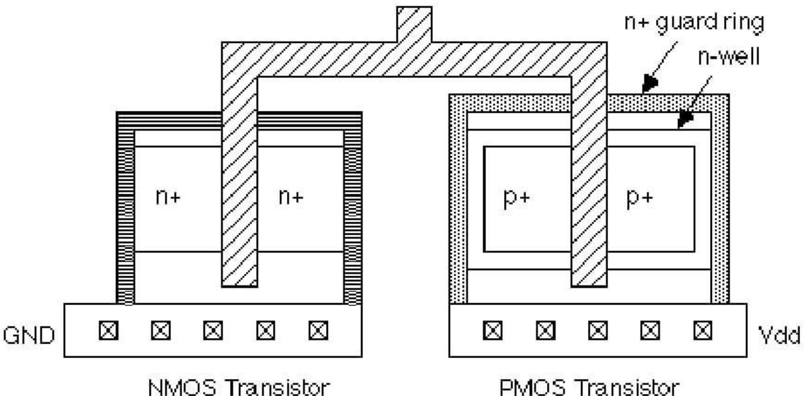
Latchup may begin when Vout drops below GND due to a noise spike or an improper circuit hookup (Vout is the base of the lateral NPN Q2). If sufficient current flows through Rsub to turn on Q2 (I Rsub > 0.7 V ), this will draw current through Rwell. If the voltage drop across Rwell is high enough, Q1 will also turn on, and a self-sustaining low resistance path between the power rails is formed. If the gains are such that b1 x b2 > 1, latchup may occur. Once latchup has begun, the only way to stop it is to reduce the current below a critical level, usually by removing power from the circuit.

The most likely place for latch up to occur is in pad drivers, where large voltage transients and large currents are present.

Preventing latch up

Fab/Design Approaches

1. Reduce the gain product b1 x b1
    - move n-well and n+ source/drain farther apart increases width of the base of Q2 and reduces gain beta2 > also reduces circuit density
    - buried n+ layer in well reduces gain of Q1
2. Reduce the well and substrate resistances, producing lower voltage drops
    - higher substrate doping level reduces Rsub
    - reduce Rwell by making low resistance contact to GND
    - guard rings around p- and/or n-well, with frequent contacts to the rings, reduces the parasitic resistances.




- Surrounding PMOS and NMOS transistors with an insulating oxide layer (trench). This breaks parasitic SCR structure.
- Latch up Protection Technology circuitry which shuts off the device when latchup is detected.

# UNIT II

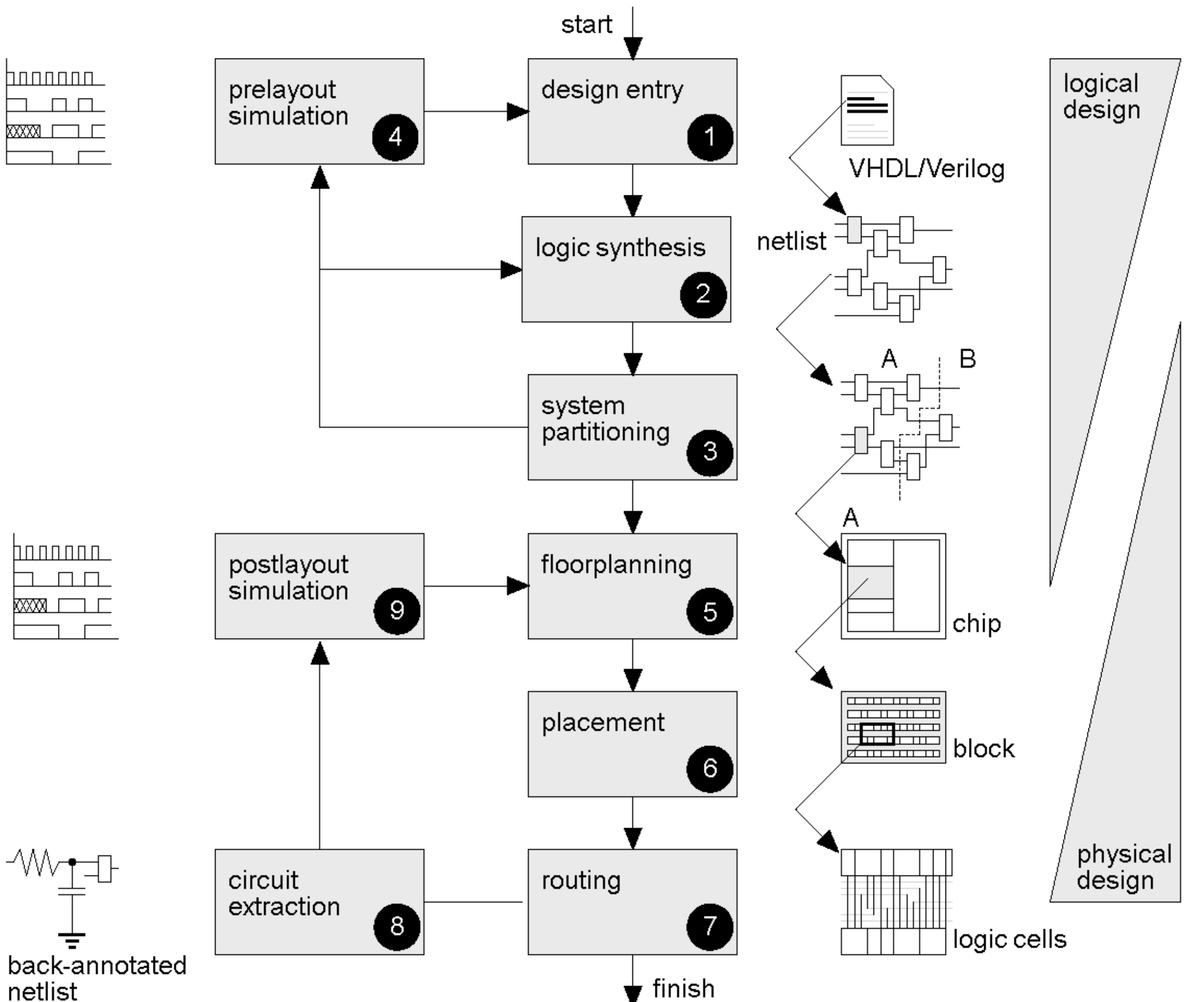
## VLSI Circuit Design Processes

- **VLSI Design Flow**
- **MOS Layers**
- **Stick Diagrams**
- **Design Rules and Layout**
- **Lamda (λ) based design rules for wires, contacts and Transistors**
- **Layout Diagrams for NMOS and CMOS Inverters and Gates**
- **Scaling of MOS circuits**

# VLSI DESIGN FLOW

A design flow is a sequence of operations that transform the IC designers' intention (usually represented in RTL format) into layout GDSII data.

A well-tuned design flow can help designers go through the chip-creation process relatively smoothly and with a decent chance of error-free implementation. And, a skilful IC implementation engineer can use the design flow creatively to shorten the design cycle, resulting in a higher likelihood that the product will catch the market window.



**Front-end design (Logical design):**

**1. Design entry** – Enter the design in to an ASIC design system using a hardware description language (HDL) or schematic entry

**2. Logic synthesis** – Generation of net list (logic cells and their connections) from HDL code. 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** - Divide a large system into ASIC-sized pieces

**4. Pre-layout simulation** - Check to see if the design functions correctly. Gate level functionality and timing details can be verified.

**Back-end design (Physical design):**

**5. Floor planning** - Arrange the blocks of the netlist on the chip

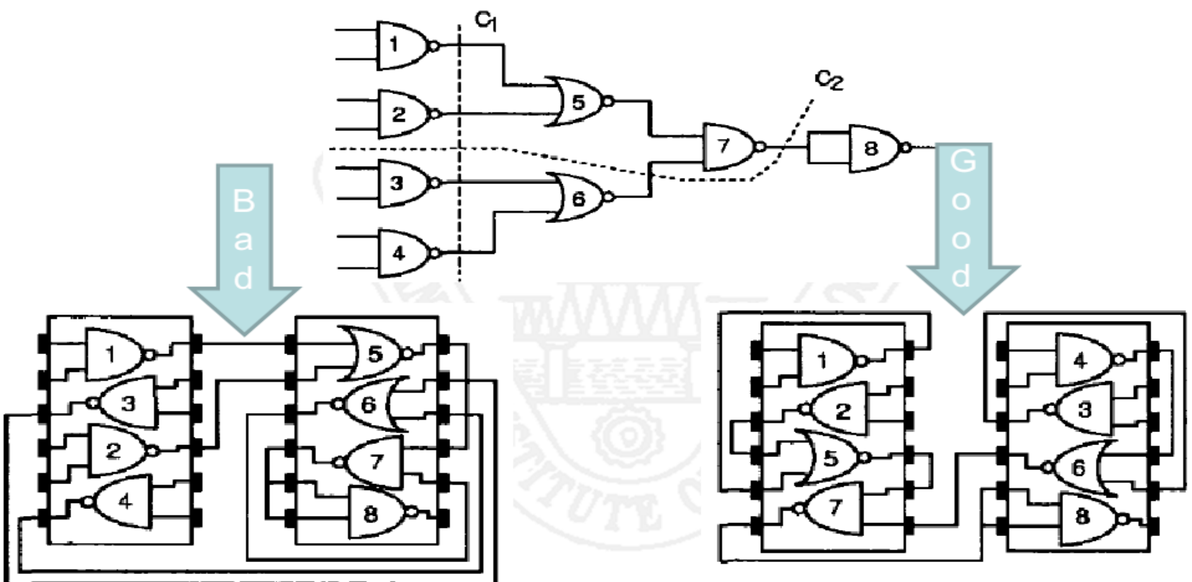
**6. Placement** - Decide the locations of cells in a block

**7. Routing** - Make the connections between cells and blocks

**8. Circuit Extraction** - Determine the resistance and capacitance of the interconnect

**9. Post-layout simulation** - Check to see the design still works with the added loads of the interconnect

**Partitioning**



## MOS LAYERS

MOS design is aimed at turning a specification into masks for processing silicon to meet the specification. We have seen that MOS circuits are formed on four basic layers

- N-diffusion
- P-diffusion
- Poly Si
- Metal

which are isolated from one another by thick or thin (thinox) silicon silicon dioxide insulating layers. The thin oxide (thinox) mask region includes n-diffusion, p-diffusion, and transistor channels. Polysilicon and thinox regions interact so that a transistor is formed where they cross one another.

## STICK DIAGRAMS

A stick diagram is a diagrammatic representation of a chip layout that helps to abstract a model for design of full layout from traditional transistor schematic. Stick diagrams are used to convey the layer information with the help of a color code.

***"A stick diagram is a cartoon of a layout."***

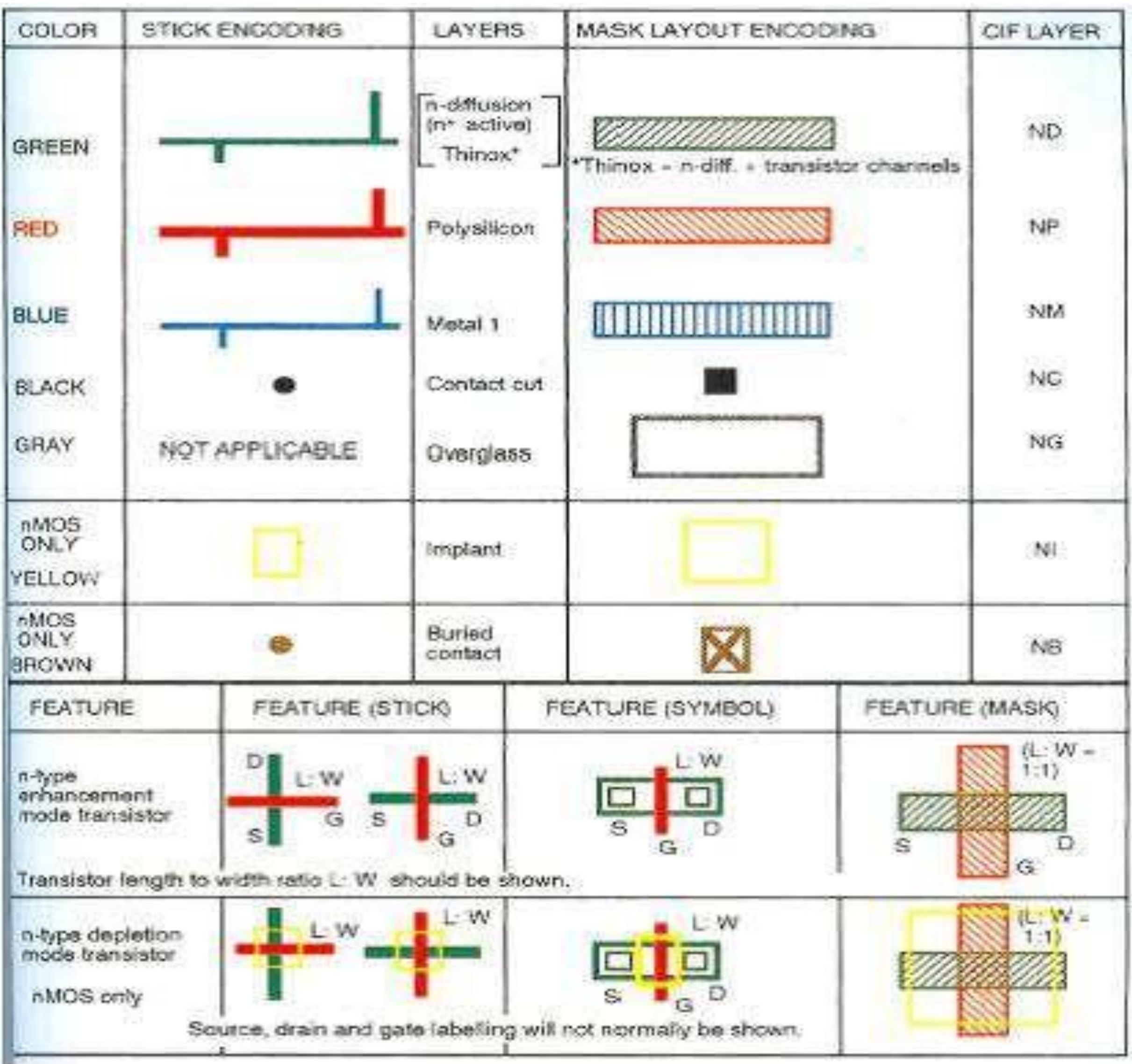
The designer draws a freehand sketch of a layout, using colored lines to represent the various process layers such as diffusion, metal and polysilicon. Where polysilicon crosses diffusion, transistors are created and where metal wires join diffusion or polysilicon, contacts are formed.

For example, in the case of nMOS design,

- Green color is used for n-diffusion
- Red for polysilicon
- Blue for metal
- Yellow for implant, and black for contact areas.

Monochrome encoding is also used in stick diagrams to represent the layer information.

## Stick Diagrams –NMOS Encoding

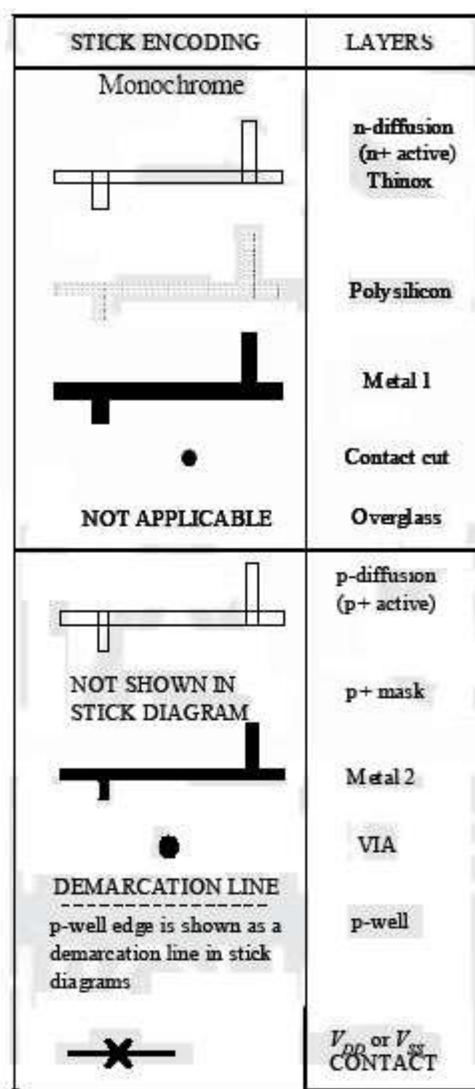


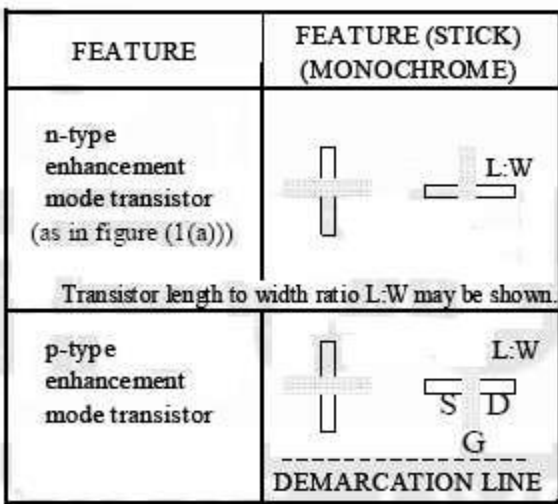
| COLOR | STICK ENCODING | LAYERS | MASK LAYOUT ENCODING | CIF LAYER |
|---|---|---|---|---|
| GREEN | | n-diffusion (n+ active) Thinox* | *Thinox = n-diff. + transistor channels | ND |
| RED | | Polysilicon | | NP |
| BLUE | | Metal 1 | | NM |
| BLACK | | Contact cut | | NC |
| GRAY | NOT APPLICABLE | Overglass | | NG |
| nMOS ONLY YELLOW | | Implant | | NI |
| nMOS ONLY BROWN | | Buried contact | | NB |

| FEATURE | FEATURE (STICK) | FEATURE (SYMBOL) | FEATURE (MASK) |
|---|---|---|---|
| n-type enhancement mode transistor | D L: W G S; L: W S D G | L: W S G D | (L: W = 1:1) S D G |
| Transistor length to width ratio L: W should be shown. | | | |
| n-type depletion mode transistor nMOS only | L: W; L: W | L: W S G D | (L: W = 1:1) |
| Source, drain and gate labelling will not normally be shown. | | | |

NMOS ENCODING

## CMOS ENCODING



## Stick Diagrams – Some Rules

**Rule 1:**

When two or more 'sticks' of the same type cross or touch each other that represents electrical contact.

**Rule 2:**

When two or more "sticks" of different type cross or touch each other there is no electrical contact. (If electrical contact is needed we have to show the connection explicitly)

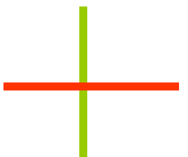
**Rule 3:**

When a poly crosses diffusion it represents a transistor.

Note: If a contact is shown then it is ***not*** a transistor.

**Rule 4:**

In CMOS a demarcation line is drawn to avoid touching of p-diff with n-diff. All PMOS must lie on one side of the line and all NMOS will have to be on the other side.



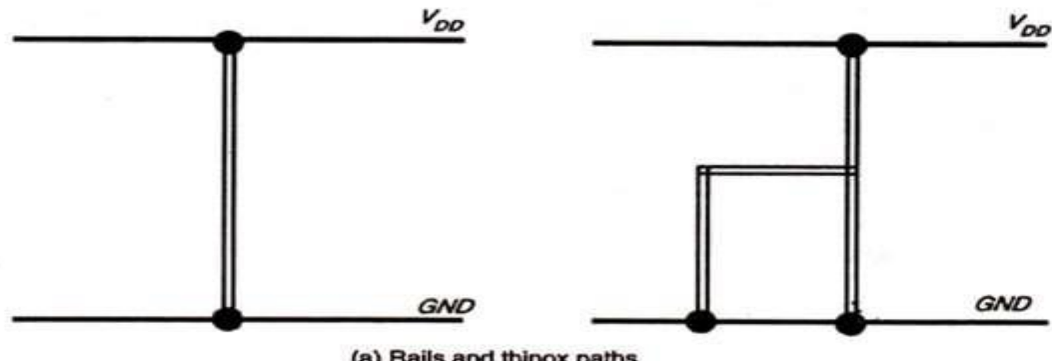
**nMOS Design Style :**

To understand the design rules for nMOS design style , let us consider a single metal, single polysilicon nMOS technology.

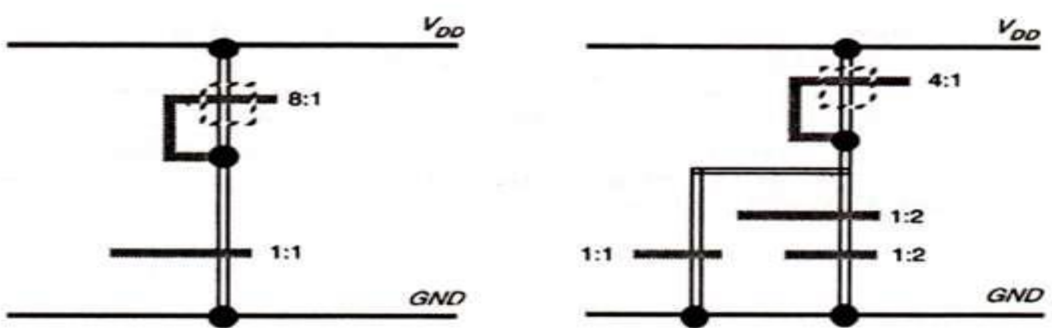
The layout of nMOS is based on the following important features.

- ✓ n-diffusion [n-diff.] and other thin oxide regions [thinox] (green) ;
- ✓ polysilicon 1 [poly.]-since there is only one polysilicon layer here (red);
- ✓ metal 1 [metal]-since we use only one metal layer here (blue);
- ✓ implant (yellow);
- ✓ contacts (black or brown [buried]).

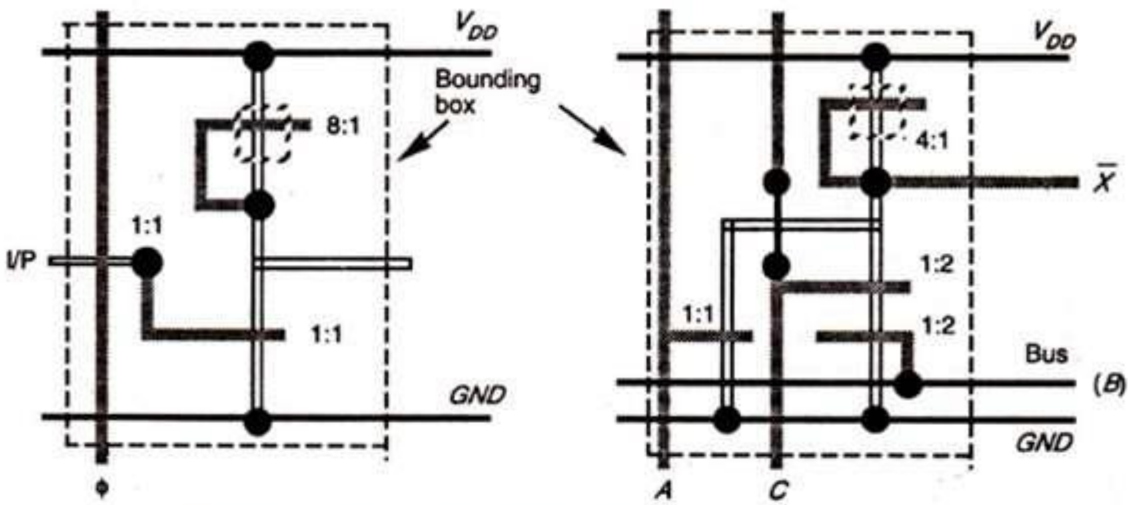
A transistor is formed wherever poly. crosses n-diff. (red over green) and all diffusion wires (interconnections) are n-type (green).When starting a layout, the first step normally taken is to draw the metal (blue) $V_{DD}$ and GND rails in parallel allowing enough space between them for the other circuit elements which will be required. Next, thinox (green) paths may be drawn between the rails for inverters and inverter based logic as shown in Fig. below. Inverters and inverter-based logic comprise a pull-up structure, usually a depletion mode transistor, connected from the output point to $V_{DD}$ and a pull down structure of enhancement mode transistors suitably interconnected between the output point and GND. This is illustrated in the Fig.(b). remembering that poly. (red) crosses thinox (green)wherever transistors are required. One should consider the implants (yellow) for depletion mode transistors and also consider the length to width (L:W) ratio for each transistor. These ratios are important particularly in nMOS and nMOS- like circuits.



(a) Rails and thinox paths



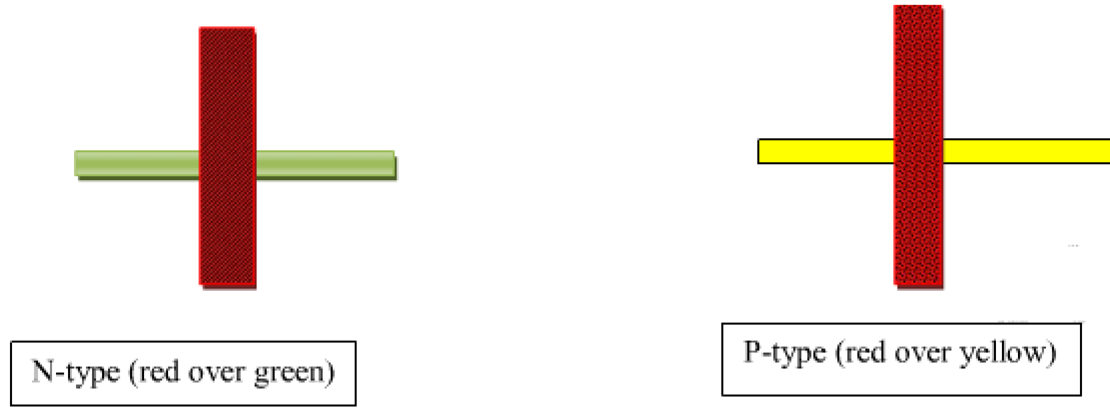
(b) Pull-up and pull-down structures (polysilicon), implants, and ratios



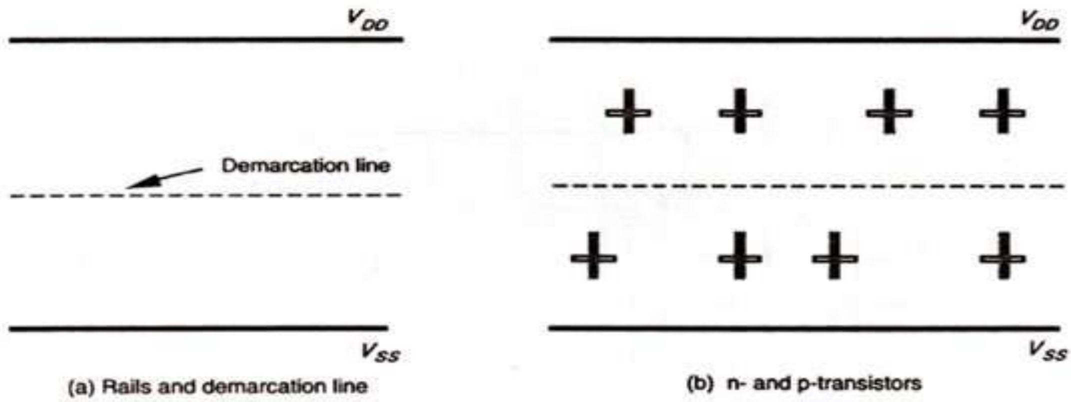
(c) Buses, control signals, interconnections, and 'leaf-cell' boundaries

**CMOS Design Style:**

The CMOS design rules are almost similar and extensions of n-MOS design rules except the Implant (yellow) and the buried contact (brown). In CMOS design Yellow is used to identify p transistors and wires, as depletion mode devices are not utilized. The two types of transistors 'n' and 'p', are separated by the demarcation line (representing the p-well boundary) above which all p-type devices are placed (transistors and wires (yellow). The n-devices (green) are consequently placed below the demarcation line and are thus located in the p-well as shown in the diagram below.




Diffusion paths must not cross the demarcation line and n-diffusion and p-diffusion wires must not join. The 'n' and 'p' features are normally joined by metal where a connection is needed. Their geometry will appear when the stick diagram is translated to a mask layout. However, one must not forget to place crosses on VDD and Vss rails to represent the substrate and p-well connection respectively. The design style is explained by taking the example the design of a single bit shift register. The design begins with the drawing of the VDD and Vss rails in parallel and in metal and the creation of an (imaginary) demarcation line in-between, as shown in Fig.below. The n-transistors are then placed below this line and thus close to Vss, while p-transistors are placed above the line and below VDD In both cases, the transistors are conveniently placed with their diffusion paths parallel to the rails (horizontal in the diagram) as shown in Fig.(b). A similar approach can be taken with transistors in symbolic form.



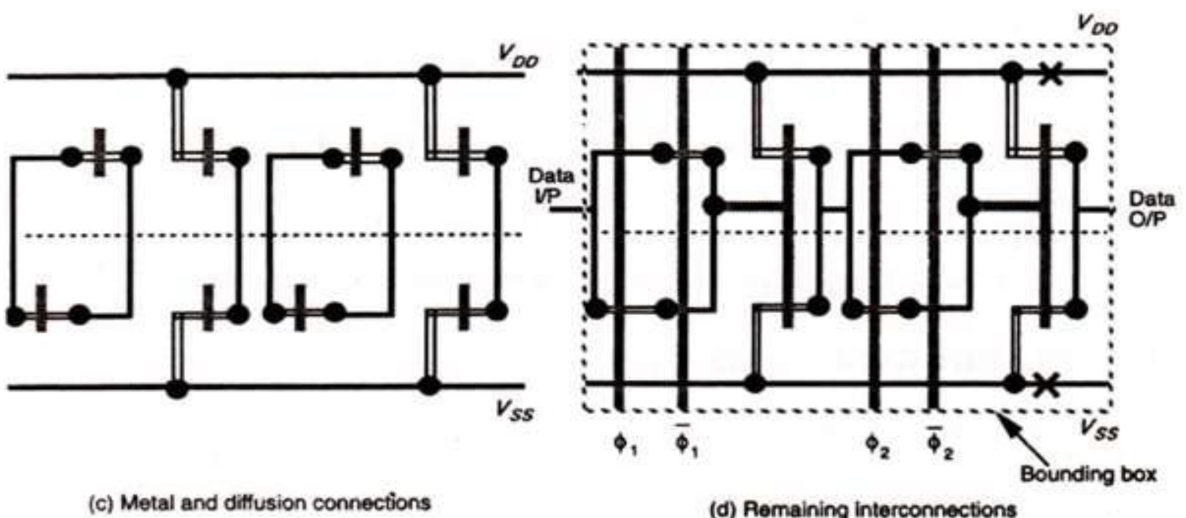




**Fig. CMOS stick layout design style (a,b,c,d)**

The n- along with the p-transistors are interconnected to the rails using the metal and connect as Shown in Fig.(d). It must be remembered that only metal and poly-silicon can cross the demarcation line but with that restriction, wires can run-in diffusion also. Finally, the remaining interconnections are made as appropriate and the control signals and data inputs are added as shown in the Fig.(d).

## Stick Diagrams:

P- Diffusion

n- Diffusion

Poly silicon

Metal 1

Contact cut

N implant

Demarcation line

Substrate contact

Buried Contact

PMOS Enhancement Transistor

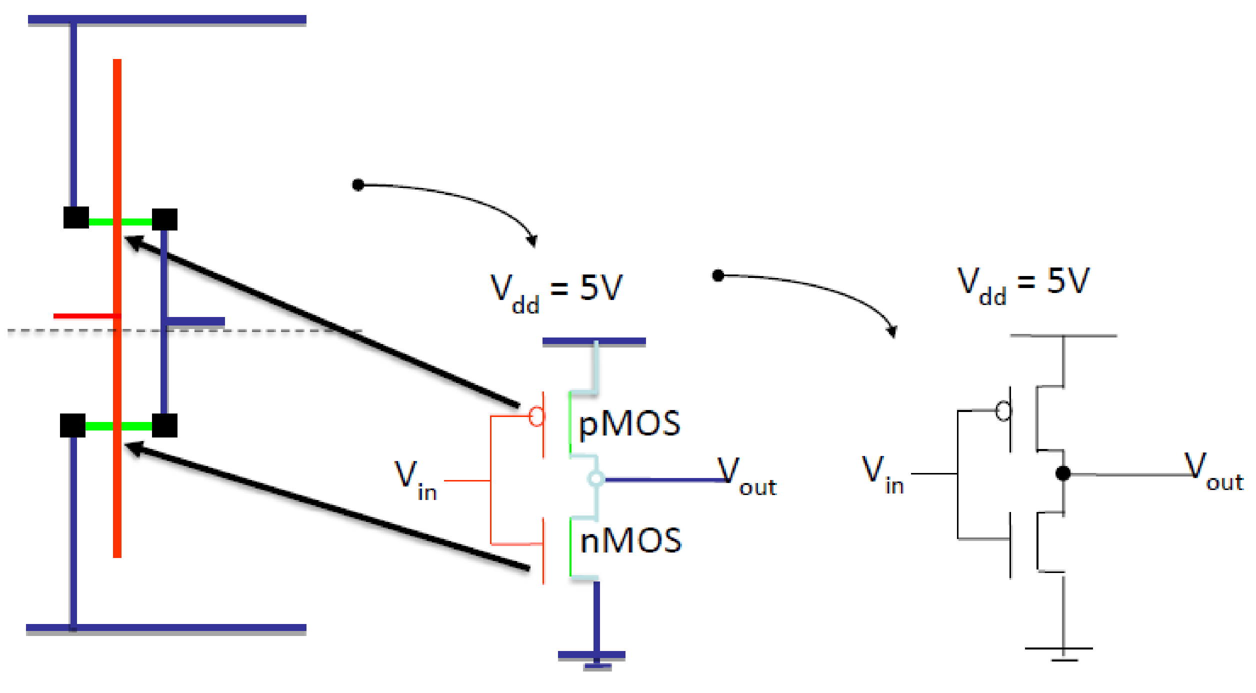
NMOS Enhancement Transistor

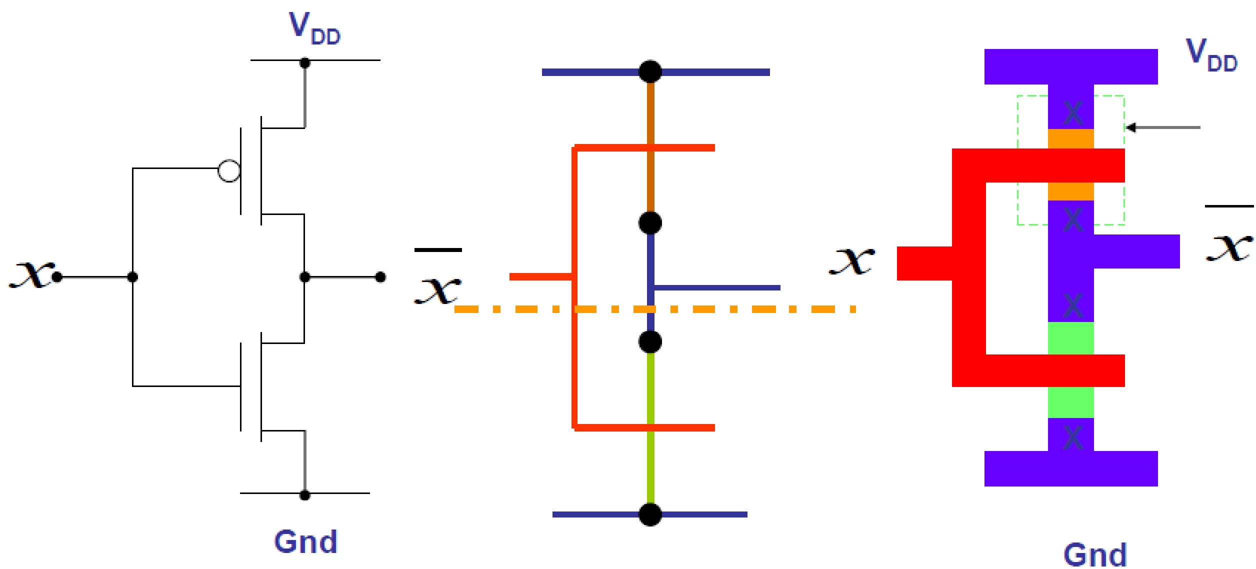
NMOS Depletion transistor

NPN Bipolar Transistor

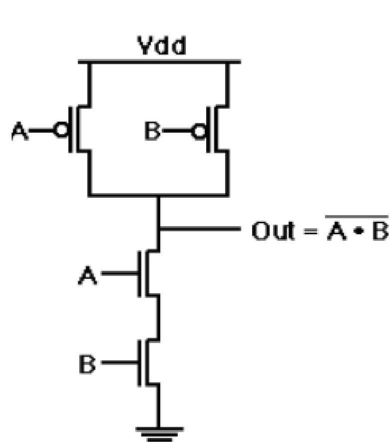
## Examples of Stick Diagrams

## CMOS Inverter

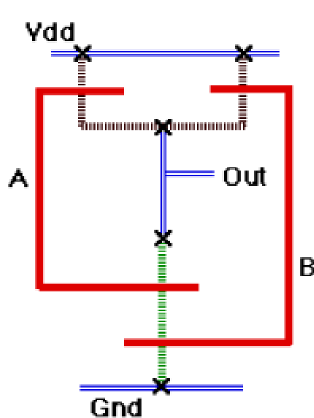


Contd….



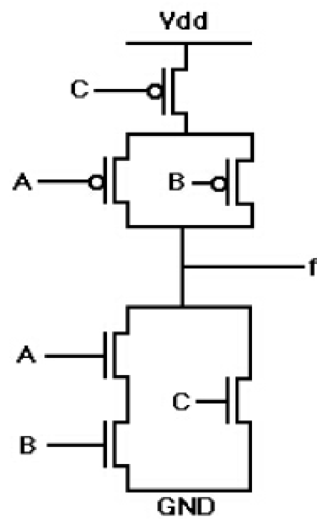

| A | B | $\overline{A \cdot B}$ |
|---|---|---|
| 0 | 0 | 1 |
| 0 | 1 | 1 |
| 1 | 0 | 1 |
| 1 | 1 | 0 |

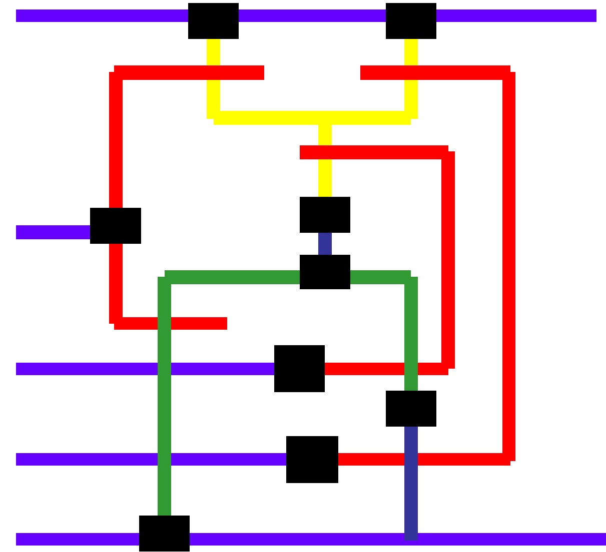



1. Pull-down: Connect to ground If A=1 AND B=1
2. Pull-up: Connect to Vdd If A=0 OR B=0

Fig. CMOS NAND gate

Example: $f = \overline{(A \cdot B) + C}$



## Design Rules and Layout

In VLSI design, as processes become more and more complex, need for the designer to understand the intricacies of the fabrication process and interpret the relations between the different photo masks is really troublesome. Therefore, a set of layout rules, also called **design rules**, has been defined. They act as an interface or communication link between the circuit designer and the process engineer during the manufacturing phase. The objective associated with layout rules is to obtain a circuit with optimum yield (functional circuits versus non-functional circuits) in as small as area possible without compromising reliability of the circuit. In addition, Design rules can be conservative or aggressive, depending on whether yield or performance is desired. Generally, they are a compromise between the two. Manufacturing processes have their inherent limitations in accuracy. So the need of design rules arises due to manufacturing problems like –

- Photo resist shrinkage, tearing.
- Variations in material deposition, temperature and oxide thickness.
- Impurities.
- Variations across a wafer.

These lead to various problems like :

- **Transistor problems:**

  Variations in threshold voltage: This may occur due to variations in oxide thickness, ion-implantation and poly layer. Changes in source/drain diffusion overlap. Variations in substrate.

- **Wiring problems:**

  Diffusion: There is variation in doping which results in variations in resistance, capacitance. Poly, metal: Variations in height, width resulting in variations in resistance, capacitance. Shorts and opens.

- **Oxide problems:**

  Variations in height.

  Lack of planarity.

- **Via problems:**

  Via may not be cut all the way through.

Undersize via has too much resistance.

Via may be too large and create short.

To reduce these problems, the design rules specify to the designer certain geometric constraints on the layout artwork so that the patterns on the processed wafers will preserve the topology and geometry of the designs. This consists of minimum-width and minimum-spacing constraints and requirements between objects on the same or different layers. Apart from following a definite set of rules, design rules also come by experience.

**Why we use design rules?**

- Interface between designer and process engineer
- Historically, the process technology referred to the length of the silicon channel between the source and drain terminals in field effect transistors.
- The sizes of other features are generally derived as a ratio of the channel length, where some may be larger than the channel size and some smaller.

For example, in a 90 nm process, the length of the channel may be 90 nm, but the width of the gate terminal may be only 50 nm.

**Semiconductor manufacturing processes**

- 10 µm — 1971
- 3 µm — 1975
- 1.5 µm — 1982
- 1 µm — 1985
- 800 nm (0.80 µm) — 1989
- 600 nm (0.60 µm) — 1994
- 350 nm (0.35 µm) — 1995
- 250 nm (0.25 µm) — 1998
- **180 nm** (0.18 µm) — 1999
- 130 nm (0.13 µm) — 2000
- 90 nm — 2002
- 65 nm — 2006
- 45 nm — 2008
- 32 nm — 2010
- 22 nm — approx. 2011
- 16 nm — approx. 2018
- 11 nm — approx. 2022

Design rules define ranges for features

Examples:

- min. wire widths to avoid breaks
- min. spacing to avoid shorts
- minimum overlaps to ensure complete overlaps
  - Measured in microns
  - Required for resolution/tolerances of masks

Fabrication processes defined by minimum channel width

- Also minimum width of poly traces
- Defines “how fast” a fabrication process is

## Types of Design Rules

The design rules primary address two issues:

1. The geometrical reproduction of features that can be reproduced by the maskmaking and lithographical process, and

2. The interaction between different layers.

There are primarily two approaches in describing the design rules.

1. Linear scaling is possible only over a limited range of dimensions.

2. Scalable design rules are conservative .This results in over dimensioned and less dense design.

3. This rule is not used in real life.

**1. Scalable Design Rules (e.g. SCMOS, λ-based design rules):**

In this approach, all rules are defined in terms of a single parameter **λ**. The rules are so chosen that a design can be easily ported over a cross section of industrial process ,making the layout portable .Scaling can be easily done by simply changing the value of.

The key disadvantages of this approach are:

2. **Absolute Design Rules (e.g. μ-based design rules ) :**

In this approach, the design rules are expressed in absolute dimensions (e.g. 0.75μm) and therefore can exploit the features of a given process to a maximum degree. Here, scaling and porting is more demanding, and has to be performed either manually or using CAD tools .Also, these rules tend to be more complex especially for deep submicron.

The fundamental unity in the definition of a set of design rules is the minimum line width .It stands for the minimum mask dimension that can be safely transferred to the semiconductor material .Even for the same minimum dimension, design rules tend to differ from company to company, and from process to process. Now, CAD tools allow designs to migrate between compatible processes.

## LAMBDA-BASED DESIGN RULES:-

- *Lambda-based* (scalable CMOS) design rules define scalable rules based on λ (which is half of the minimum channel length)
  - classes of MOSIS SCMOS rules: SUBMICRON, DEEPSUBMICRON
- Stick diagram is a draft of real layout, it serves as an abstract view between the schematic and layout.
- Circuit designer in general want tighter, smaller layouts for improved performance and decreased silicon area.
- On the other hand, the process engineer wants design rules that result in a controllable and reproducible process.
- Generally we find there has to be a compromise for a competitive circuit to be produced at a reasonable cost.
- All widths, spacing, and distances are written in the form
- $\lambda = 0.5$ X minimum drawn transistor length
- Design rules based on single parameter, λ
- Simple for the designer
- Wide acceptance
- Provide feature size independent way of setting out mask
- If design rules are obeyed, masks will produce working circuits
- Minimum feature size is defined as 2 λ
- Used to preserve topological features on a chip
- Prevents shorting, opens, contacts from slipping out of area to be contacted

# LAMBDA BSED RULES

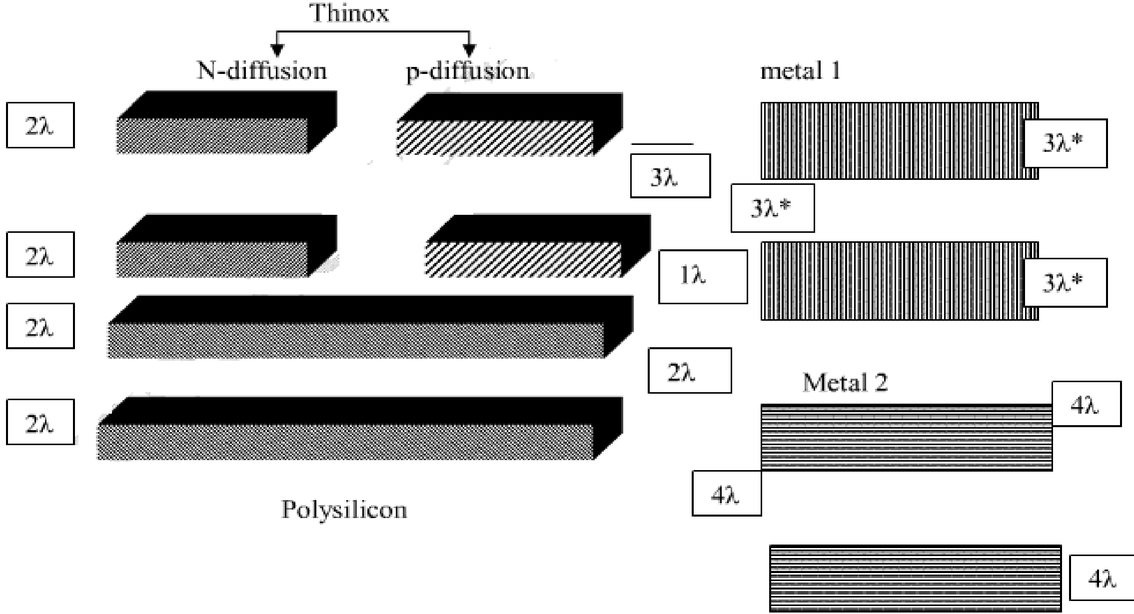
MINIMUM WIDTH AND SPACING RULES

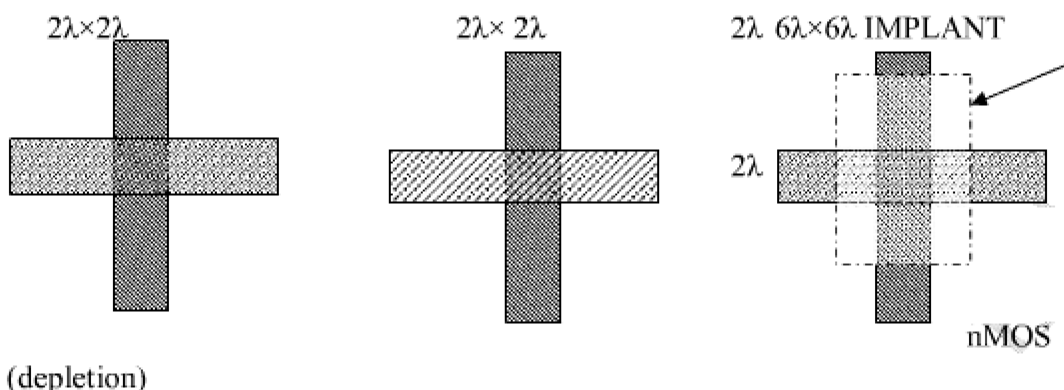
| LAYER | TYPE OF RULE | VALUE |
|---|---|---|
| POLY | Minimum Width<br>Minimum Spacing | 2λ<br>2λ |
| N/P DIFFUSION | Minimum Width<br>Minimum Spacing | 3λ<br>3λ |
| N-WELL | Minimum Width<br>Minimum Spacing | 3λ<br>3λ |
| P-WELL | Minimum Width<br>Minimum Spacing | 3λ<br>3λ |
| METAL1 | Minimum Width<br>Minimum Spacing | 3λ<br>3λ |

## DESIGN RULES FOR WIRES (nMOS and CMOS)

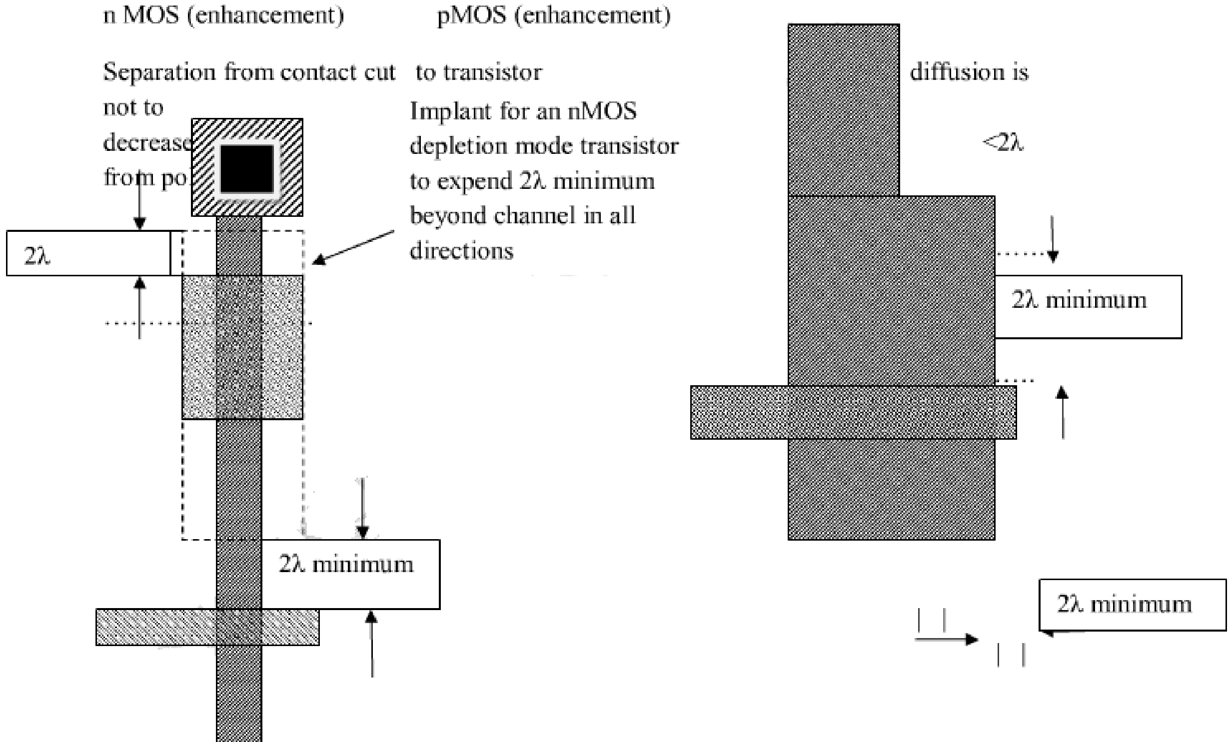
Design rules and layout methodology based on the concept of λ provide a process and feature size independent way of setting out mask dimensions to scale. All paths in layers are dimensioned in λ units and subsequently λ can be allocated an appropriate value compatible with the feature size of the fabrication process.

Minimum width specified) minimum separation (where



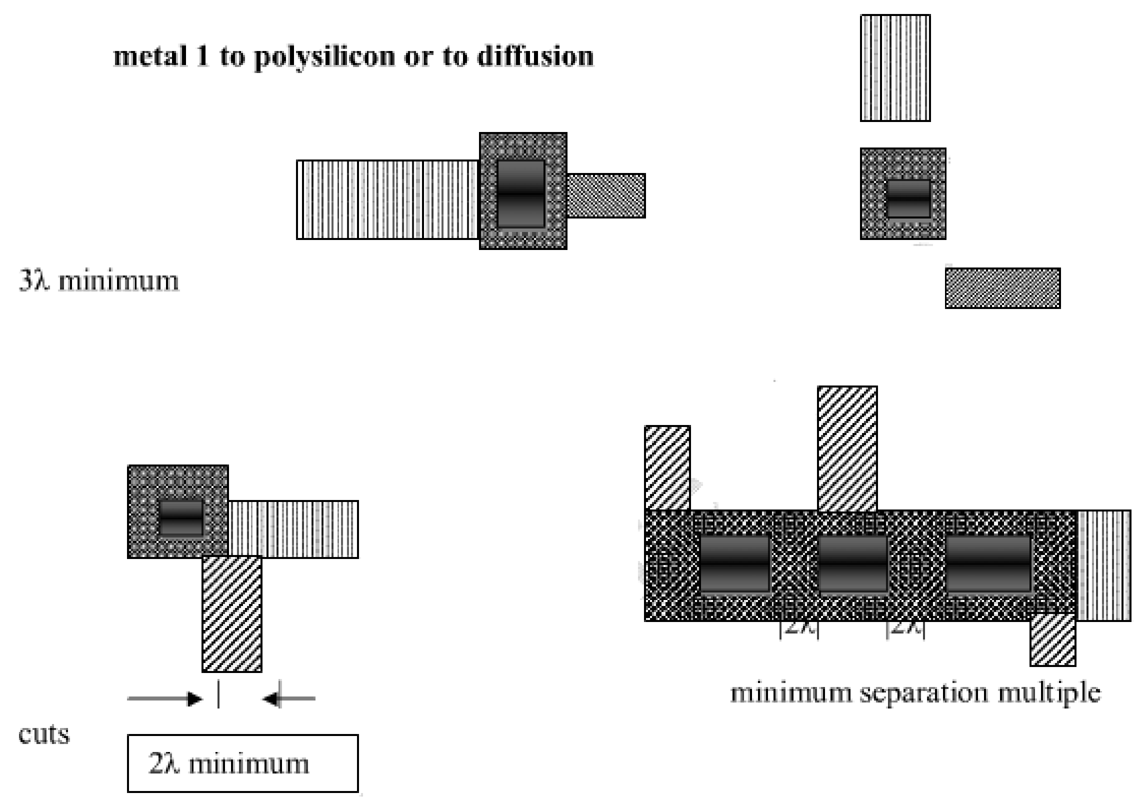



(depletion)



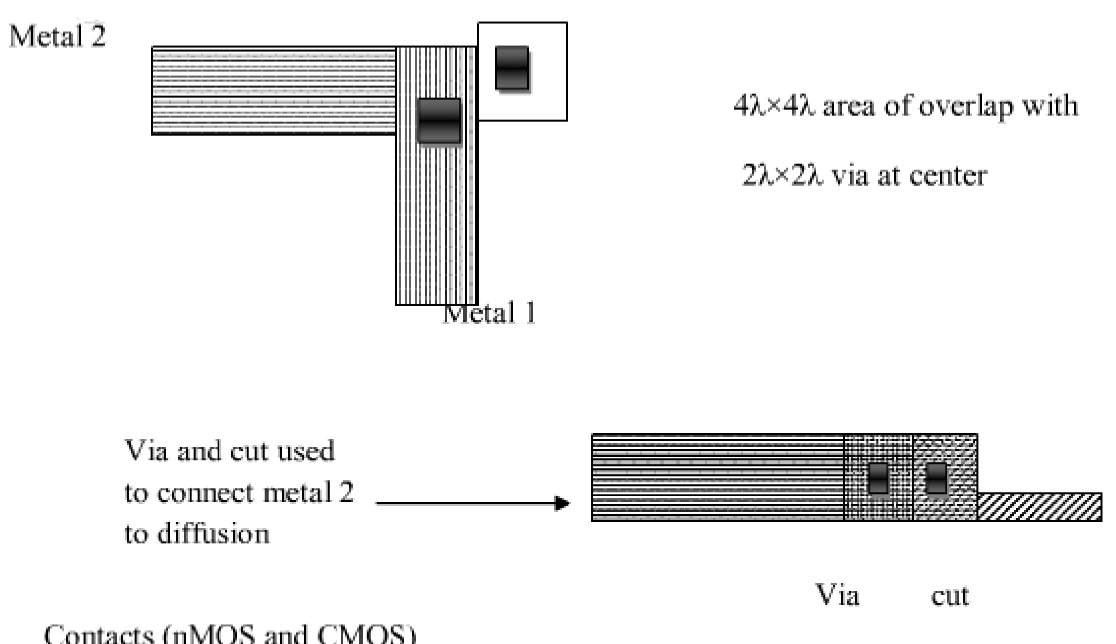

polysilicon to extend a minimum of 2λ beyond diffusion boundaries (width constant)

Separation from implant to another transistor




2λ×2λ cut centered on 4λ×4λ superimposed area of layers to be joined in all cases

2λ minimum separation (if other spacing's allowed)



## CONTACT CUTS

When making contacts between poly-silicon and diffusion in nMOS circuits it should be remembered that there are three possible approaches--poly. to metal then metal to diff., or aburied contact poly. to diff. , or a butting contact (poly. to diff. using metal). Among the three the latter two, the buried contact is the most widely used, because of advantage in space and a reliable contact. At one time butting contacts were widely used , but now a days they are superseded by buried contacts.

In CMOS designs, poly. to diff. contacts are always made via metal. A simple process is followed for making connections between metal and either of the other two layers (as in Fig.a), The 2λ. x 2λ. contact cut indicates an area in which the oxide is to be removed down to the underlying polysilicon or diffusion surface. When deposition of the metal layer takes place the metal is deposited through the contact cut areas onto the underlying area so that contact is made between the layers.

The process is more complex for connecting diffusion to poly-silicon using the butting contact approach (Fig.b), In effect, a 2λ. x 2λ contact cut is made down to each of the layers to be joined. The layers are butted together in such a way that these two contact cuts become contiguous. Since the poly-silicon and diffusion outlines overlap and thin oxide under poly silicon acts as a mask in the diffusion process, the poly-silicon and diffusion layers are also butted together. The contact between the two butting layers is then made by a metal overlay as shown in the Fig.

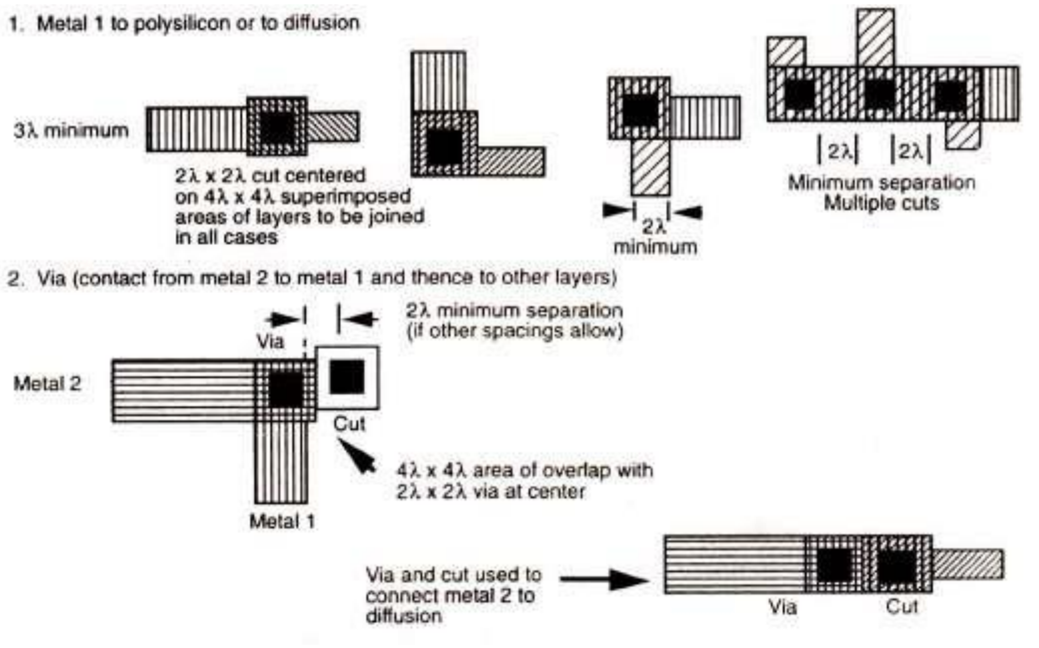



Fig.(a) . n-MOS & C-MOS Contacts

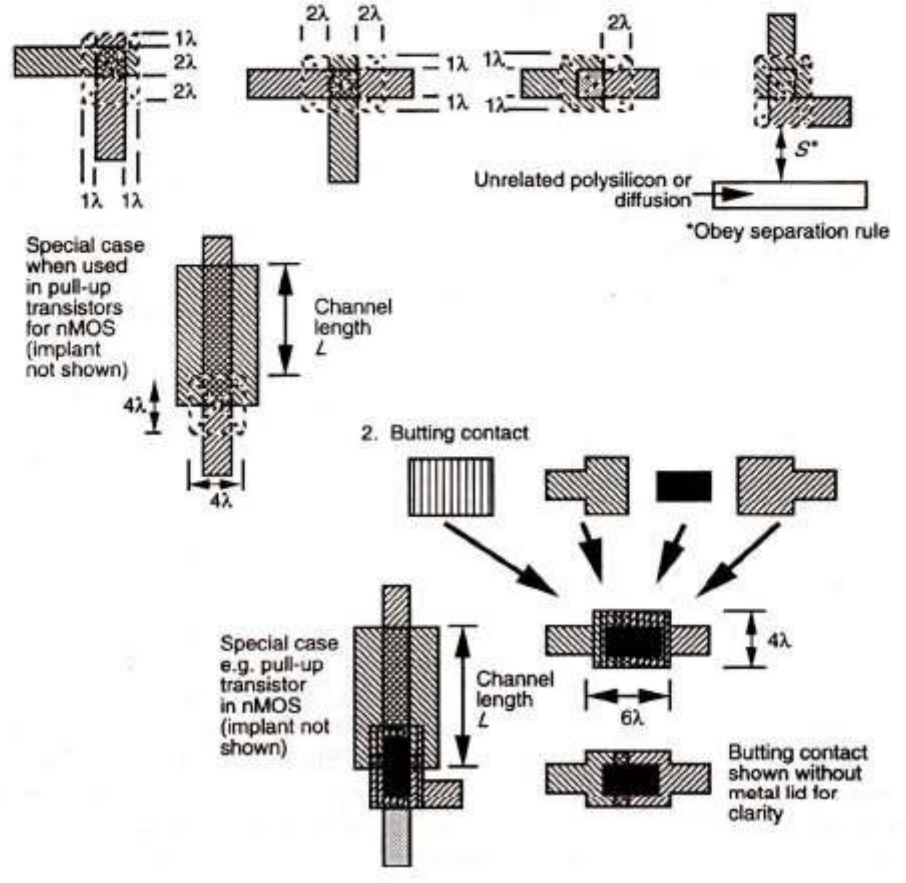



Fig.(b). Contacts poly-silicon to diffusion

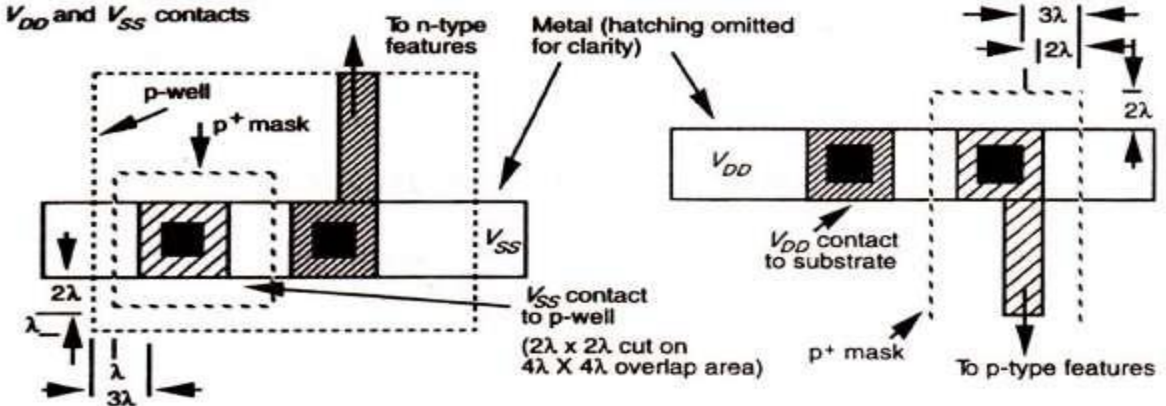
In buried contact basically, layers are joined over a 2λ. x 2λ. area with the buried contact cut extending by 1λ, in all directions around the contact area except that the contact cut extension is increased to 2λ. in diffusion paths leaving the contact area. This helps to avoid the formation of unwanted transistors. So this buried contact approach is simpler when compared to others. The, poly-silicon is deposited directly on the underlying crystalline wafer. When diffusion takes place, impurities will diffuse into the poly-silicon as well as into the diffusion region within the contact area. Thus a satisfactory connection between poly-silicon and diffusion is ensured. Buried contacts can be smaller in area than their butting contact counterparts and, since they use no metal layer, they are subject to fewer design rule restrictions in a layout.

## Other design rules

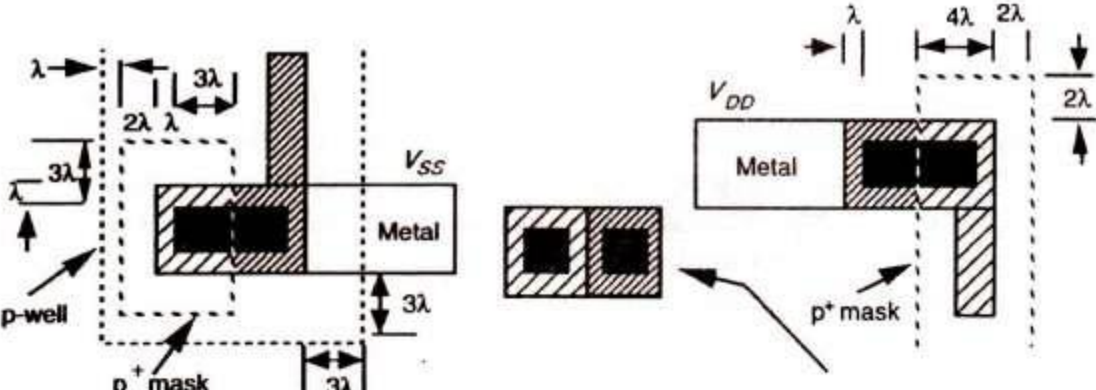
- Double Metal MOS process Rules
- CMOS fabrication is much more complex than nMOS fabrication
- 2 um Double metal, Double poly. CMOS/BiCMOS Rules
- 1.2um Double Metal single poly.CMOS rules

**CMOS Lambda-based Design Rules:**

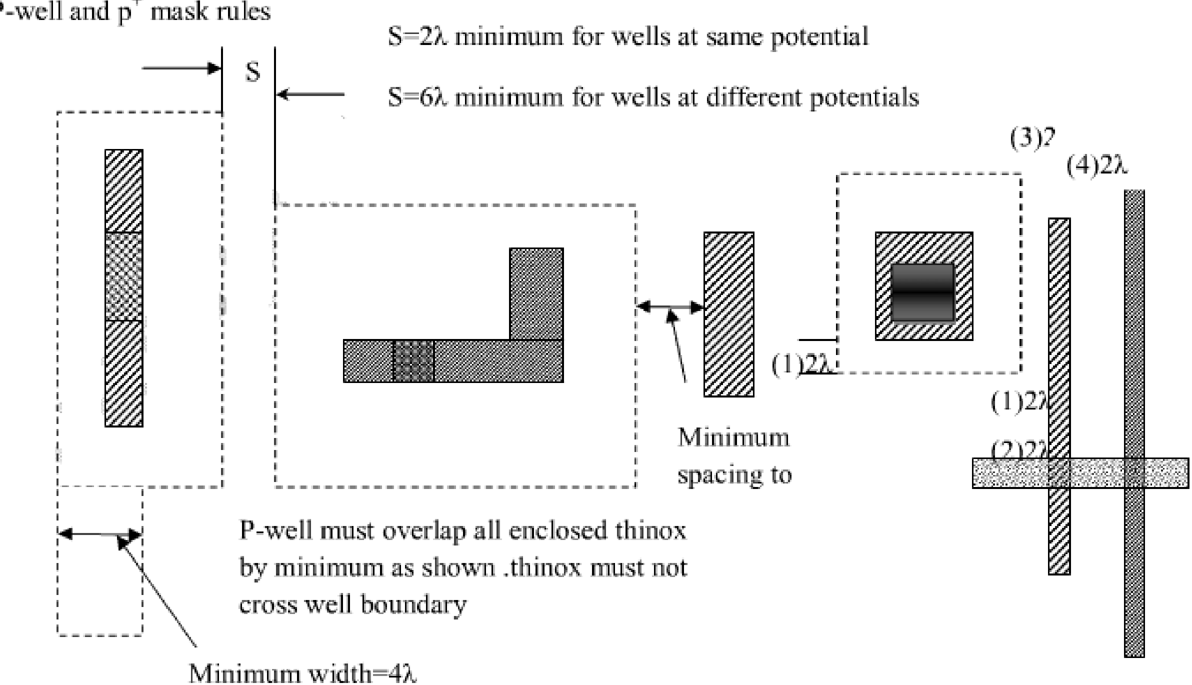
The CMOS fabrication process is more complex than nMOS fabrication. In a CMOS process, there are nearly 100 actual set of industrial design rules. The additional rules are concerned with those features unique to p-well CMOS, such as the p-well and p+ mask and the special 'substrate contacts. The p-well rules are shown in the diagram below




In the diagram above each of the arrangements can be merged into single split contacts.




From the above diagram it is also clear that split contacts may also be made with separate cuts.



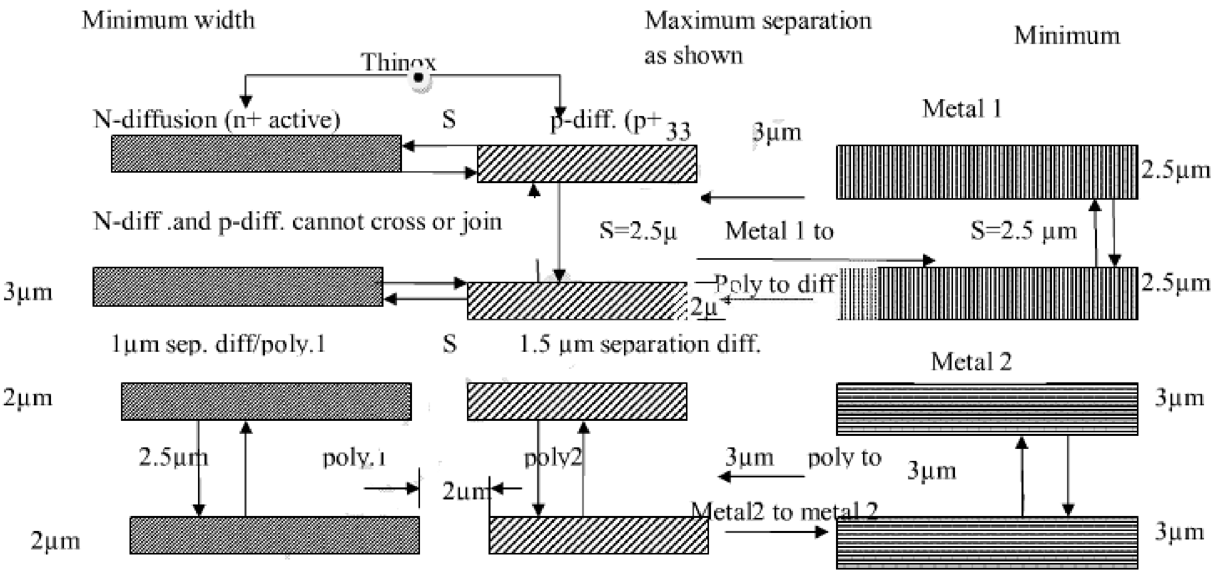

The CMOS rules are designed based on the extensions of the Mead and Conway concepts and also by excluding the butting and buried contacts the new rules for CMOS design are formed. These rules for CMOS design are implemented in the above diagrams.

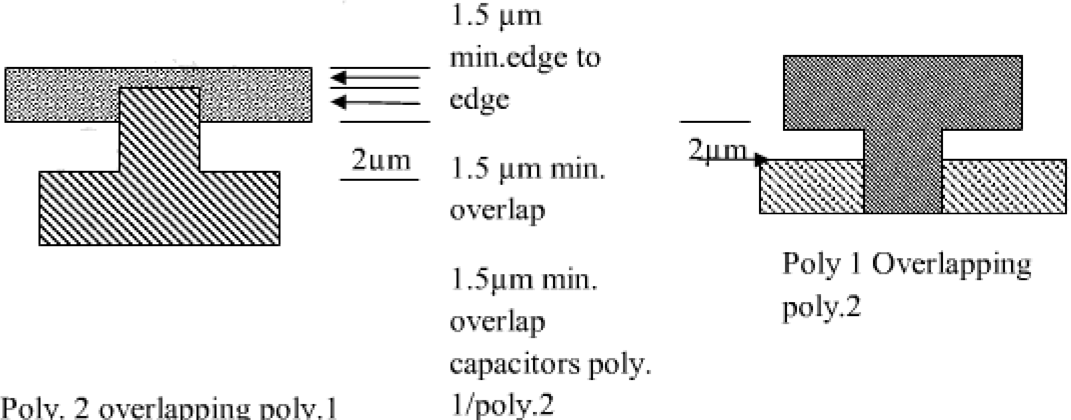
## µM CMOS Design rules

The encoding is compatible with that already described where as following extension are made: n-well → brown

Poly 1 → red; poly 2 → orange; diff (n-active) → green; p Diff (p-active) → yellow.

For BiCMOS the following are added: buried $n^+$ sub collector- pale green; p-base--pink.




Poly 1 Overlapping poly.2

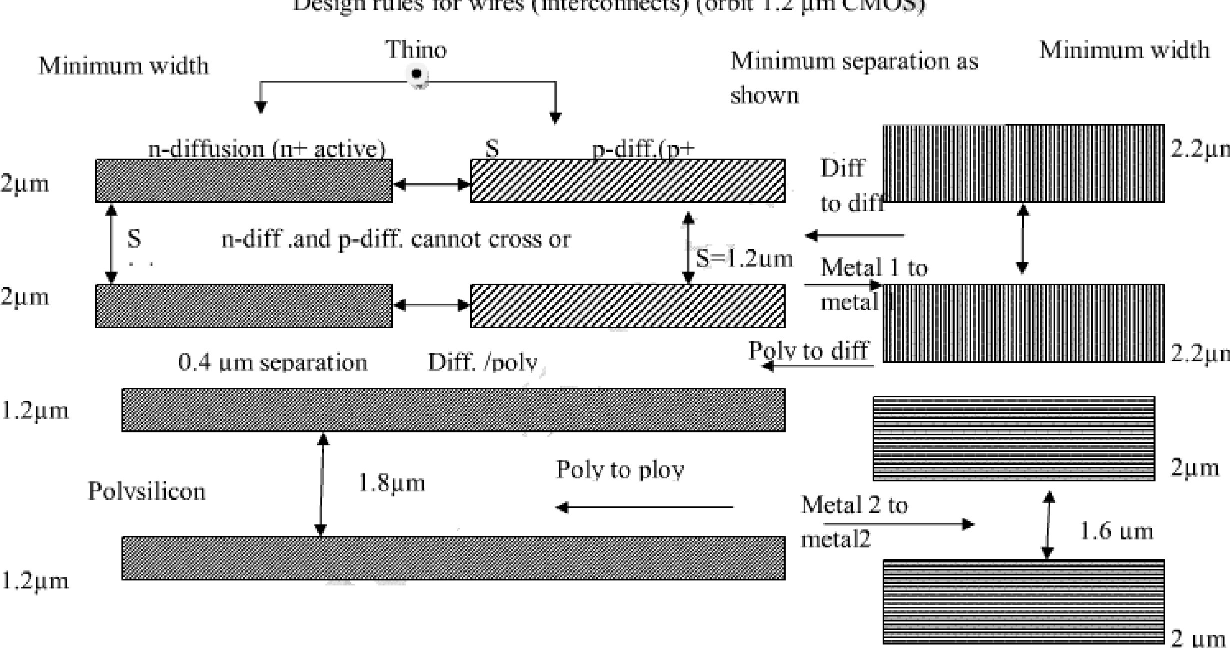
Poly. 2 overlapping poly.1

AVOID COINCIDENT EDGES WHERE METAL 1 AND METAL 2 RUNS FOLLOW THE SAME PATH FOR >25µm LENGTH (UNDER LAP METAL 1

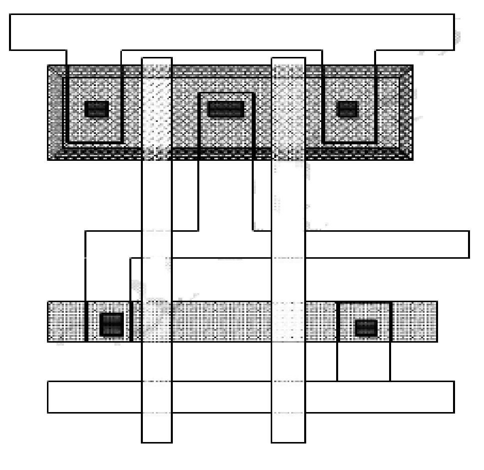
Design rules for wires (interconnects) (orbit 2µm CMOS)

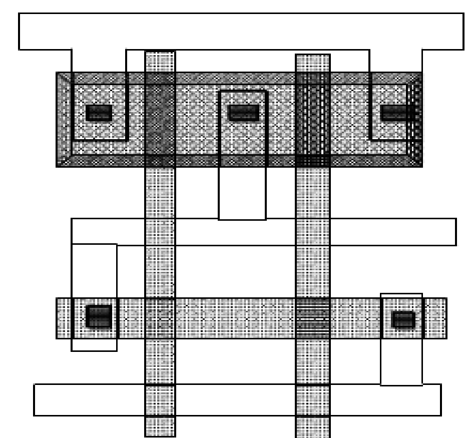
**2µm DOUBLE METAL, SINGLE POLY CMOS RULES**

The orbit ™ 1.2µm rules provide improved feature size. A separate set of micro based design rules accompany them

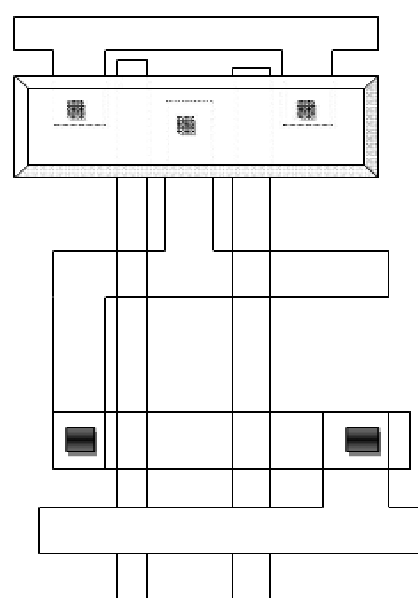


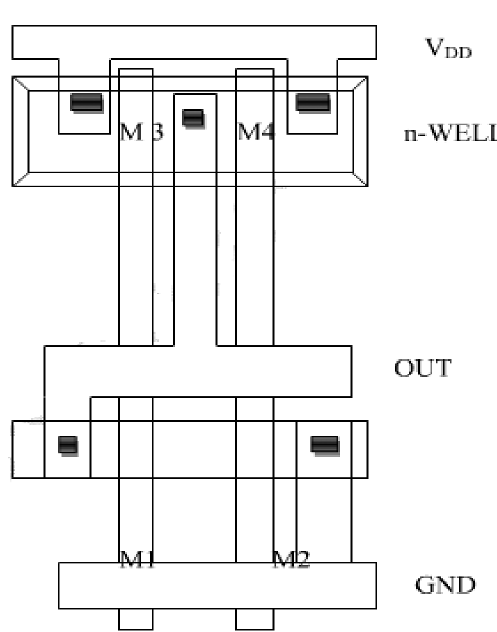

Avoid coincident edges where metal 1 and metal2 runs follow the same path for>25µm length (under lap metal 1 edges by 0.8 µm).





**N-WELL AND ACTIVE AREA MASKs AND . ..** **POLY MASK -> DEFINE NMOS**
**..........................................................................PMOS TRANSISTORS**






**Metal mask for $V_{DD}$, GND and output connections** $V_A$ $V_B$

**METAL –DIFFUSION CONSTANT MASK**

## Layout Diagrams for NMOS and CMOS Inverters and Gates

### Layer Types

- p-substrate
- n-well
- n+
- p+
- Gate oxide
- Gate (polysilicon)
- Field Oxide
  - Insulated glass
  - Provide electrical isolation

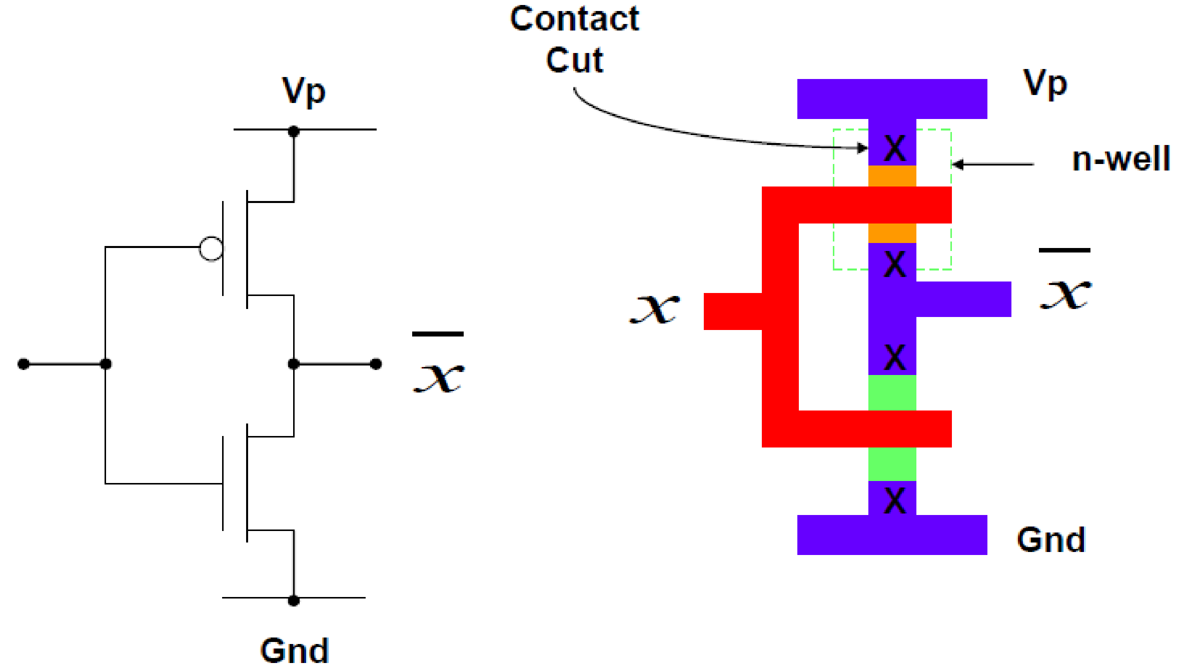
## Basic Gate Design

Both the power supply and ground are routed using the Metal layer

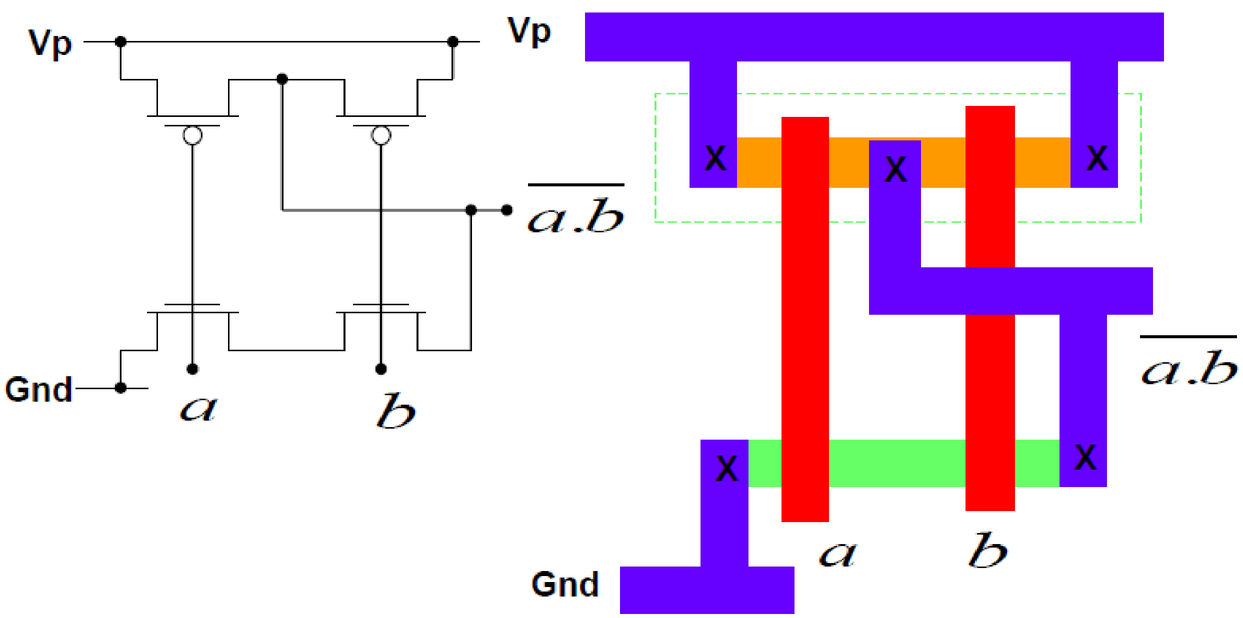
n+ and p+ regions are denoted using the same fill pattern. The only difference is the n-well

Contacts are needed from Metal to n+ or p+

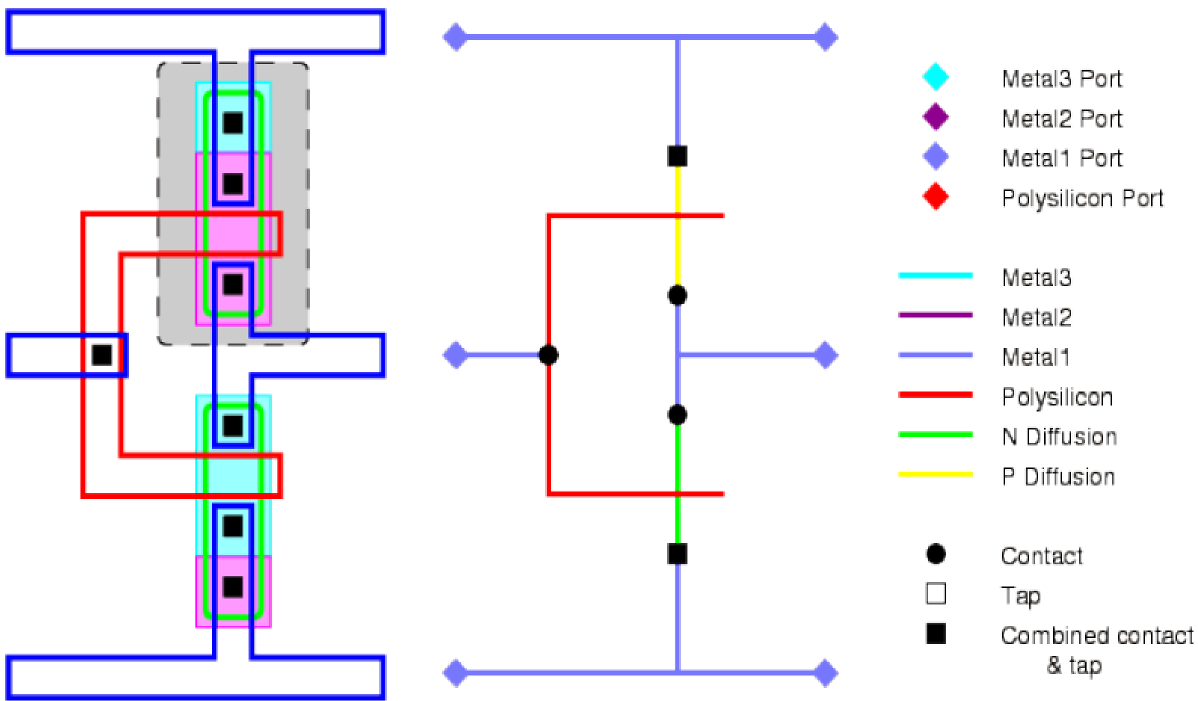
# The CMOS NOT Gate



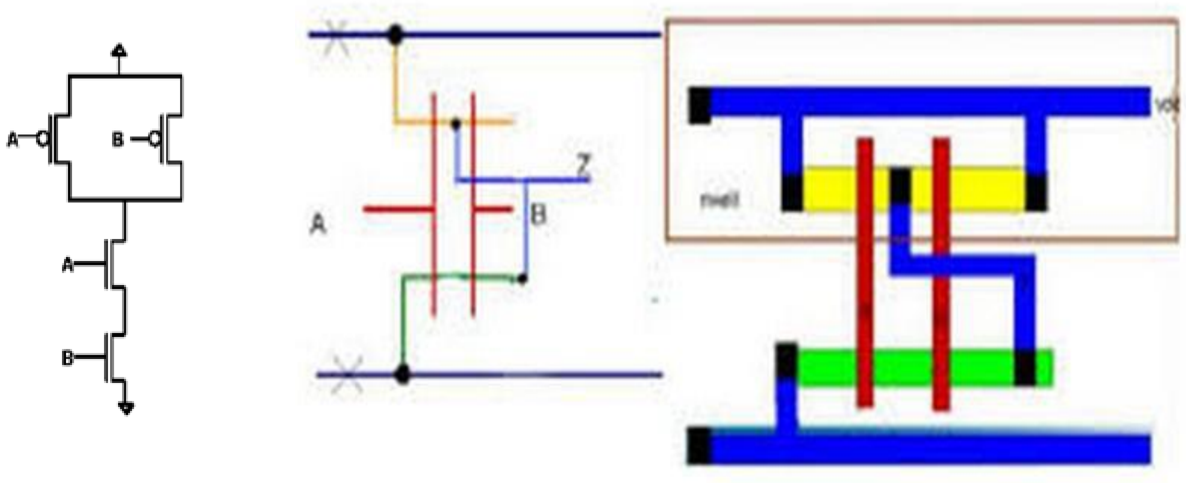

# The CMOS NAND Gate



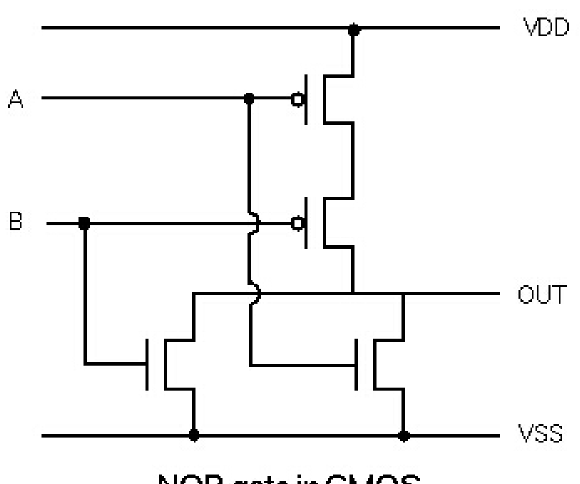
**Layout & Stick Diagram of CMOS Inverter**



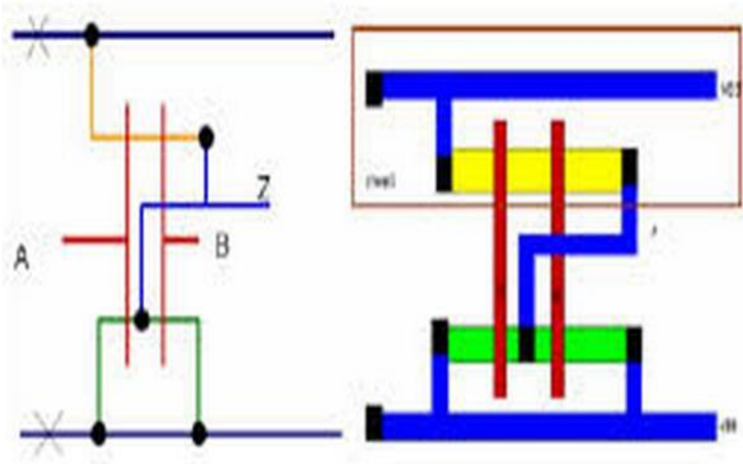

## 2 input NAND gate



## 2 input NOR gate




NOR gate in CMOS

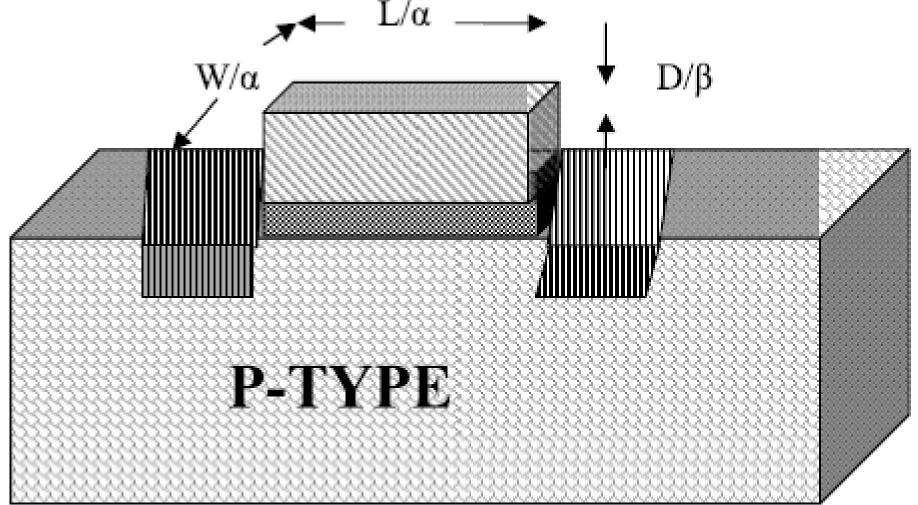



## Scaling of MOS circuits

Scaling means to reduce the feature size and to achieve higher packing density of circuitry on a chip, Many figures of merit such as minimum feature size, number of gates on one chip, power dissipation, maximum operational frequency, die size, production cost can be improved by shrinking the dimensions of transistors, interconnections and the separation between features, and by adjusting the doping levels and supply voltages.

## SCALING MODELS AND SCALING FACTORS:

The most commonly used models are the constant electric field scaling models and the constant voltage scaling model. One more model called as combined voltage and dimension scaling model is presented recently. The following figure indicates the device dimensions and substrate doping level which are associated with the scaling of a transistor.




Two scaling factors 1/α,1/β are used. 1/β is chosen as the scaling factor for supply voltage $V_{DD}$ and gate oxide thickness D, and 1/α is used for all other linear dimensions, both vertical and horizontal to chip surface.

## SCALING FACTORS FOR DEVICE PARAMETERS:

### GATE AREA AG :

$$Ag = L.W$$

Where L and W are the channel length and width respectively, both are scaled by 1/α. So Ag is scaled by $1/\alpha^2$

### GATE CAPACITANCE PER UNIT ARE $C_O$ OR $C_{OX}$:

$$C_o = E_{OX}/D$$

Where $E_{OX}$ is the permittivity of the gate oxide (thinox) $(=E_{ins}.E_O)$ and D is the gate oxide thickness which is scaled by 1/β

Thus $C_O$ is scaled by $1/1/\beta=\beta$

### GATE CAPACITANCE $C_g$:

$$C_g = C_O\ L.W$$

Thus $C_g$ is scaled by $\beta.1/\alpha^2=\beta/\alpha^2$

## PARASITIC CAPACITANCE $C_X$:

$C_X$ is proportional to $A_X/d$.

Where d is the depletion width around source or drain which is scaled by $1/\alpha$ and $A_X$ is the area of depletion region around source or drain which is scaled by $1/\alpha_2.1/1/\alpha = 1/\alpha$

## CARRIER DENSITY IN CHANNEL $Q_{on}$

$$Q_{on}=C_o.V_{gs}$$

Where Qon is the average charge per unit area in the channel in the 'on' state. $C_o$ is scaled by $\beta$ and $V_{gs}$ is scaled by $1/\beta$.

Thus $Q_{on}$ is scaled by 1.

## CHANNEL RESISTANCE $R_{on}$

$$R_{on}=L/W.Q_{on}.\mu$$

Where $\mu$ is the carrier mobility in the channel and is assumed constant. Thus $R_{on}$ is scaled by $1/\alpha.1/1/\alpha=1$.

## GATE DELAY $T_d$

**Td is proportional to $R_{on}.C_g$.**

Thus $T_d$ is scaled by $\beta^2/\alpha^4$

## MAXIMUM OPERATING FREQUENCY $F_O$:

$$F_o = W/L * \mu C_O V_{DD}/Cg$$

Or $f_o$ is inversely proportional to delay $T_d$. Thus $f_o$ is scaled by $1/\beta/\alpha^2 = \alpha^2/\beta$

## SATURATION CURRENT $I_{DSS}$:

$$I_{dss}=C_{o\mu}/2.W/L.\ (V_{gs}-V_t)^2$$

Nothing that both $V_{gs}$ and $V_t$ are scaled by $1/\beta$,we have $I_{dss}$ is scaled by $\beta(1/\beta)^2=1/\beta$.

**CURRENT DENSITY J:**

$$J=I_{des}/A$$

Where A is the cross sectional area of the channel in the 'on' state which is scaled by $1/\alpha^2$

So, J is scaled by $1/\beta/1/\alpha^2=\alpha^2/\beta$.

## SWITCHING ENERGY PER GATE $E_g$:

$$Eg=C_g/2.(V_{DD})^2$$

So Eg is scaled by $\beta/\alpha^2.1/\beta^2=1/\alpha^2\beta$

## POWER DISSIPATION PER GATE $P_g$:

$P_g$ comprise two components such that

$$P_g=P_{gs}+P_d$$

Where the static component

$$P_{gs}= (V_{DD})^2/R_{on}$$

And the dynamic component

$$P_{gd}=E_g f_o$$

It will be seen that both $P_{gs}$ and $P_{gd}$ are scaled by $1/\beta^2$

## POWER DISSIPATION PER UNIT AREA:

$$P_a = P_g / A_g$$

So Pa is scaled by $1/\beta^2/1/\alpha^2 = \alpha^2/\beta^2$

## POWER-SPEED PRODUCT $P_T$:

$$P_T = P_g . T_d$$

So PT is scaled by $1/\beta^2 . \beta/\alpha^2 = 1/\alpha^2\beta$

## Limitations of Scaling:

Scaling may cause a problem which prevents further miniaturization.

## Substrate doping: -

The built-in (junction) potential $V_B$, is small compared with $V_{DD}$.

(a) Substrate doping scaling factors:

As the channel length of a MOS transistor is reduced, the depletion region widths must also be scaled down to prevent the source and drain depletion regions

$N_B$ is thus maintained at a satisfactory level in the channel region and thus problem is reduced. But depletion width d and built in potential $V_B$ will impose limitations on scaling.

We have $E_{max} = 2V/d$

Where $E_{max}$ is the maximum electric field induced in one-sided step junction

When $N_B$ is increased by α and if $V_\alpha=0$, then $V_\beta$ is increased by ln α and d is decreased by $\sqrt{\ln \alpha/\alpha}$.

There E is increased by inverse of this factor and reaches $E_{crit}$

## Limits of miniaturization

The minimum size of transistor is determined by both process technology and the physics of the device itself.
Transistor size is defined in terms of channel length L. L can be decreased as long as there is no punch through i.e. The depletion region around source should not come closer to that around the drain. So L must be at least 2d from meeting. Depletion region width d for the junctions is given by

$$D=\sqrt{2E_{si}E_oV/qN_B}$$

Where

$E_{si}$= relative permittivity of silicon (~12)

$E_O$= permittivity of free space (=8.85x10^-14)

V=effective voltage across the junction

$V = V_{a+}V_B$

q=electron charge

$N_B$=doping level of substrate.

$V_a$= (maximum value =$V_{DD}$)=applied voltage

$V_B$=built-in (junction) potential

And $V_B = KT/q. \ln (N_B N_D/n_i^2)$

Where $N_D$ is the source or drain doping and $n_i$ is the intrinsic carrier concentration in silicon.

## Depletion width

When $N_B$ is increased, the depletion width decreases and $V_t$ increases which is not desirable.

We have $\mathbf{V_{drift} = \mu E}$

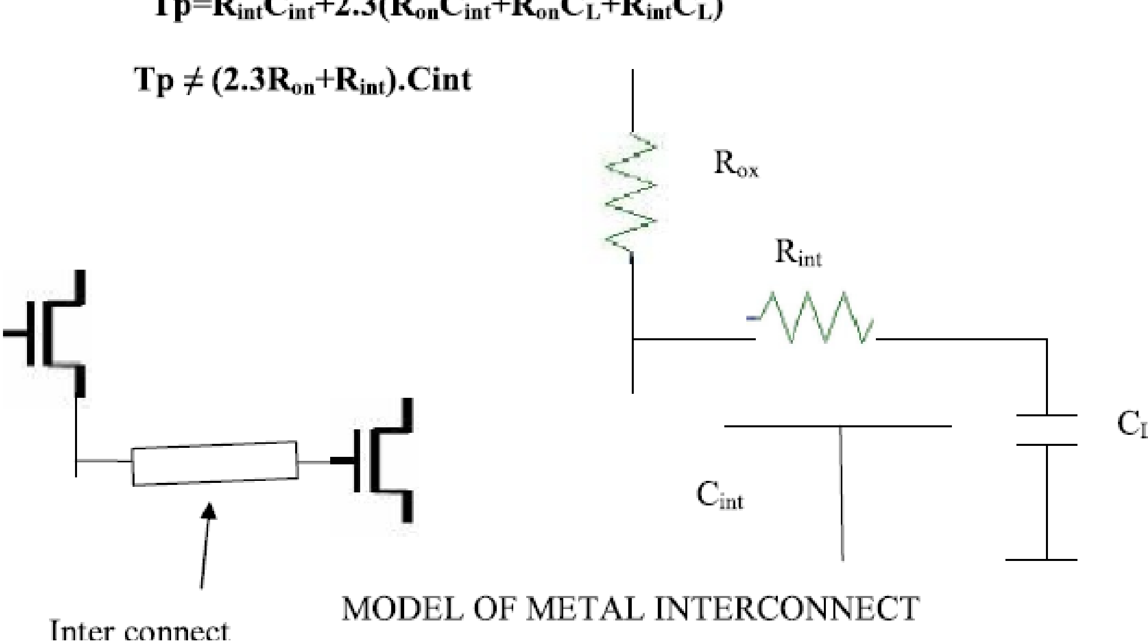
$V_{drift}$ is the carrier drift velocity and **L=2d**

Transit time $\boldsymbol{\tau} = \mathbf{L/V_{drift} = 2d/\mu E}$

## Limits due to interconnect and contact resistance

Since the width, thickness and spacing are scaled by 1/α, cross-section area must be scaled by $1/\alpha^2$ .Thus R is increased by α and I is scaled by 1/α. so IR drop remains constant. Thus driving capability and noise margins are degraded.

The propagation delay $T_p$ along a single aluminum interconnect can be calculated from the following equation




Now

$$R_{int}=\rho L/HW$$

$$C_{int} = E_{ox}[1.15W/t_{ox}+2.28(H/t_{ox})^{0.222}]L$$

Where $R_{on}$ is the ON resistance of the transistor.

$R_{int}$ is the resistance of the interconnect

$C_{int}$ is the capacitance of interconnect

$t_{ox}$ is the thickness of dielectric oxide.

$\rho$ is the resistivity of interconnect L,W,H are the length, width and height of the interconnect.

## Assignment questions:

1. Draw the circuit diagram; stick diagram and layout for CMOS inverter.
2. Explain about the various layout design rules.
3. Draw the static CMOS logic circuit for the following expression
    1. i) Y= (ABCD)′ ii) Y= [D(A+BC)]′
4. Explain in detail about the scaling concept in VLSI circuit Design.
5. Draw the Layout Diagrams for NAND Gate using nMOS..
6. Explain λ-based Design Rules in VLSI circuit Design.
7. Draw the Layout Diagrams for CMOS Inverter.
8. Discuss about the stick diagrams and their corresponding mask layout examples
9. Draw the stick diagram of p-well CMOS inverter and explain the process.
10. Explain about the 2 μm CMOS Design rules and discuss with a layout example.
11. Draw and explain the layout for CMOS 2-input NAND gate.
12. Draw the flow chart of VLSI Design flow and explain the operation of each step in detail.
13. Draw the stick diagram for three input AND gate.
14. What is the purpose of design rule? What is the purpose of stick diagram? What are the different approaches for describing the design rule? Give three approaches for making contacts between poly silicon and discussion in NMOS circuit.

**MOS and CMOS Circuit Design Process:**

MOS and CMOS circuit design process involves the concepts such as:

- MOS Layers
- Stick Diagrams
- Lambda based design rules and layout diagrams
- Basic circuit concepts such as: sheet resistance, area capacitance and delay calculation

**MOS Layers:**

MOS circuits are formed by three layers i.e. diffusion ( n or p diffusion layer), polysilicon and metal, which are isolated from one another by thick or thin (thinox) silicon dioxide insulating layers.

- The thin oxide region includes n- diffusion, p- diffusion and transistor channels. Polysilicon and thinox regions interact so that a transistor is formed where they cross one another.
- Layers may be deliberately joined together where contacts are formed.
- The basic MOS transistor properties can be modified by the use of an implant within the thinox region.

The MOS design is aimed at turning a specification into masks for processing silicon to meet the specification.

**Stick Diagrams and Layout Diagrams:**

Stick diagrams are used to convey layer information and topology through the use of color code and using these stick diagrams mask layouts can be easily designed. The color code for various layers are:

1. Green for n- diffusion
2. Red for polysilicon
3. Blue for metal
4. Yellow for implant or for p- diffusion
5. Black for contact areas

- The layout of stick diagrams faithfully reflects the topology of the actual layout in silicon and the stick diagrams are relatively easily turned into mask layouts.
- As known that the mask layout produced during design will be compatible with the fabrication processes, a set of design rules are set out for layouts so that, if obeyed, the rules will produce layouts which will work in practice.

**Mask Layout/ Layout/ Layout Diagram** represent an integrated circuit in terms of planar geometric shapes which corresponds to the pattern of the metal, oxide or semiconductor layers that make up the components of the integrated circuit. The dimensions of each layer and the separation between the layers in a layout are parameterized by $\lambda$.

## Lambda based design rules for wires (nMOS and CMOS):

### Layer width:

| Layer | Minimum Width |
|---|---|
| n- diffusion | 2λ |
| p- diffusion | 2λ |
| Polysilicon | 2λ |
| Metal 1 | 3λ |
| Metal 2 | 4λ |

### Separation between the layers:

| Layers | Minimum Separation |
|---|---|
| n- diffusion and n- diffusion | 3λ |
| p- diffusion and p- diffusion | 3λ |
| Polysilicon and polysilicon | 2λ |
| n-diffusion and polysilicon | 1λ |
| p-diffusion and polysilicon | 1λ |
| Metal 1 and metal 1 | 3λ |
| Metal 2 and metal 2 | 4λ |

## Basic circuit concepts:

Basic circuit concepts help us to calculate the actual resistance, capacitance, delay values associated with the transistors and their circuit wiring and parasitic.

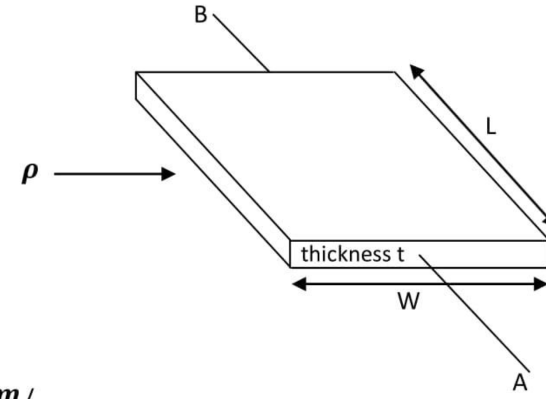
## Sheet Resistance $R_s$:

Sheet resistance is defined as the ratio of resistivity ρ and thickness t for a sheet/ slab.

Consider a uniform slab of conducting material ρ of width W, thickness t, and length L between the faces A and B, then the value of resistance of the slab (sheet) is given as,

$$R_{AB} = \frac{\rho L}{A}\ ohm$$

$$R_{AB} = \frac{\rho L}{t\,W}\ ohm$$




Where:

A = t W = area of cross section of the slab

If L = W, i.e. square of resistive material, then

$$R_{AB} = \frac{\rho}{t} = R_s\ in\ {}^{ohm}/_{square}$$

Where:

$R_s$ = sheet resistance or ohm per square

For a MOSFET transistor

$$R = ZR_s = \frac{L}{W} R_s = 4 \times 10^4 ohm$$

Where:

Z = L/ W

- It is to be noted that $R_s$ is completely independent of the area of the square.

The typical sheet resistances $R_s$ for various MOS layers are (considering different technologies)

| Layer | Sheet Resistance $R_s$ | | |
|---|---|---|---|
| | 5 µm Technology | 2 µm Technology | 1.2 µm Technology |
| Metal | 0.03 | 0.04 | 0.04 |
| n- channel transistor/ pMOS transistor | $1\times10^4$ | $2\times10^4$ | $2\times10^4$ |
| p- channel transistor/ pMOS transistor | $2.5\times10^4$ | $4.5\times10^4$ | $4.5\times10^4$ |
| Diffusion | 10- 50 | 10- 50 | 10- 50 |
| Silicide | 2- 4 | 2- 4 | 2- 4 |
| Polysilicon | 15- 100 | 15- 100 | 15- 100 |

**Area Capacitance:**

In MOS transistor conducting layers are separated from the substrate and each other by insulating (dielectric) layers, and thus parallel plate capacitive effects are present and are allowed.

For any layer, knowing the dielectric (silicon dioxide) thickness, we can calculate area capacitance as,

$$C = \frac{\varepsilon_0\, \varepsilon_{ins}\, A}{D} = \frac{k\, A}{D}\, farads$$

Where:

D = thickness of silicon dioxide

k = dielectric constant

A = Area of plates

$\varepsilon_{ins} = relative\ permittivity\ of\ silicon\ dioxide$

$\varepsilon_0 = relative\ permittivity\ of\ free\ space = 8.85 \times 10^{-14} F/\,cm$

- Normally area capacitances are given in pF/ μm$^2$ (where μm = micron = $10^{-6}$ meter = $10^{-4}$ cm). The appropriate figure may be calculated as:

$$C\left(\frac{pF}{\mu m^2}\right)=\frac{\varepsilon_0\,\varepsilon_{ins}}{D}\frac{F}{cm^2}\times\frac{10^{12}pF}{F}\times\frac{cm^2}{10^8\mu m^2}$$

The typical area capacitance values for 5μm MOS circuits are:

| Capacitance | Value in pF/ μm$^2$ | Relative value |
|---|---|---|
| Gate to Channel | $4\times10^{-4}$ | 1 |
| Diffusion to substrate | $1\times10^{-4}$ | 0.25 |
| Polysilicon to substrate | $0.4\times10^{-4}$ | 0.1 |
| Metal 1 to substrate | $0.3\times10^{-4}$ | 0.075 |
| Metal 2 to substrate | $0.2\times10^{-4}$ | 0.05 |
| Metal 2 to metal 1 | $0.4\times10^{-4}$ | 0.1 |
| Metal 2 to polysilicon | $0.3\times10^{-4}$ | 0.075 |

Note: Relative value = Specified value / gate to channel value

- **Standard unit of capacitance "$\square C_g$":**

The standard unit of capacitance is denoted by $\square C_g$ and is defined as the gate- to- channel capacitance of the minimum size (2λ × 2λ) MOS transistor.

- The standard unit of capacitance has provided a convenience to various MOS technologies but which can be used in calculations without associating it with an absolute value.
- $\square C_g$ can be evaluated for any MOS technology.

For example for a 5 μm MOS circuit with λ = 2.5 μm:

Gate area = 5 μm × 5 μm = 25 μm$^2$

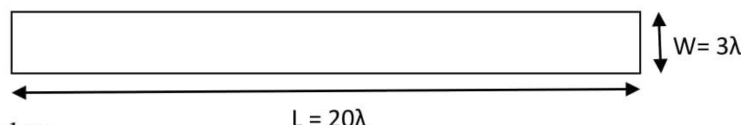
Capacitance value = 4 × $10^{-4}$ pF/ μm$^2$ (using table)

Standard Capacitance $\square C_g$= 25 μm$^2$ × 4 × $10^{-4}$ pF/ μm$^2$ = .01 pF

- **Some Area Capacitance calculations:**

Here the calculation of capacitance values may now be done by the ratio between the area of interest and the area of the standard gate (2λ × 2λ) and multiplying this ratio by the appropriate relative C value (using the table). The product will give the required capacitance in $\square C_g$ units.

Let's calculate the capacitance of a simple area of length 20λ and width 3λ respectively.




Now we will calculate:

1. Relative Area

$$Relative\ Area = \frac{L \times W}{2\lambda \times 2\lambda} = \frac{20\lambda \times 3\lambda}{2\lambda \times 2\lambda} = 15$$

2. Capacitance to substrate considering the area in metal.

$$Capacitance\ to\ substrate = relative\ area \times relative\ C\ value\ (from\ table)$$
$$Capacitance\ to\ substrate = 15 \times 0.075 \square C_g$$

3. Capacitance to substrate considering the area in polysilicon.

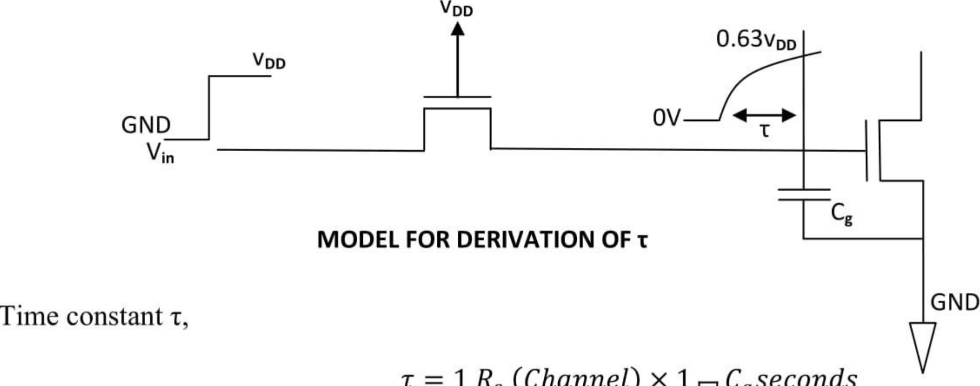
$$Capacitance\ to\ substrate = relative\ area \times relative\ C\ value\ (from\ table)$$
$$Capacitance\ to\ substrate = 15 \times 0.1 \square C_g = 1.5 \square C_g$$

4. Capacitance to substrate considering the area in diffusion.

$$Capacitance\ to\ substrate = relative\ area \times relative\ C\ value\ (from\ table)$$
$$Capacitance\ to\ substrate = 15 \times 0.25 \square C_g = 3.75 \square C_g$$

**<u>Delay Calculation/ The delay unit</u> (τ):**

Considering the case of one standard gate capacitance being charged through one square of channel resistance (from 2λ by 2λ nMOS pass transistor).




Time constant τ,

$$\tau = 1\ R_s\ (Channel) \times 1 \square C_g seconds$$

The time constant given as above can be evaluated for 5 μm technology so that,

$$Theoretical\ \tau = 10^4 ohm \times 0.01pF = 0.1\ nsec$$

In practice there are circuit wiring and parasitic capacitances, so τ is increased by a factor 2 or 3 so that for a 5 μm circuit (λ = 2.5 μm),

$$\tau = 0.2 \rightarrow 0.3\ nsec\ is\ typical\ figure$$

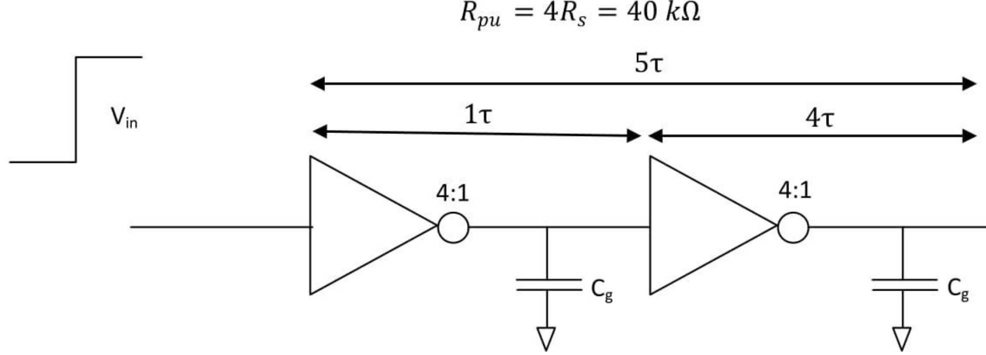
It is to be noted that τ thus obtained is not much different from transit time $\tau_{sd}$, which is given as,

$$\tau_{sd} = \frac{L^2}{\mu_n V_{ds}}$$

## Inverter Delays:

Considering a basic 4:1 ratio nMOS inverter in order to achieve the 4:1 $Z_{pu}$ to $Z_{pd}$ ratio, $R_{pu}$ will be 4 $R_{pd}$, and if $R_{pd}$ is contributed by the minimum size transistor then, clearly, the resistance value associated with $R_{pu}$ is such,

$$R_{pu} = 4R_s = 40\ k\Omega$$




- The $R_{pd}$ value is 1 $R_s$ = 10 kΩ so that the delay associated with the inverter will depend on whether it is being turned on or off and if considering the pair of cascaded inverters, then delay over the pair will be constant irrespective of the sense of the logic level transition of the input to the first. (Assuming $\tau = 0.3$ nsec and making no extra allowances fro wiring capacitance). We have an overall delay of $\tau + 4\tau = 5\tau$.

In general terms the delay through a pair of similar nMOS inverters is

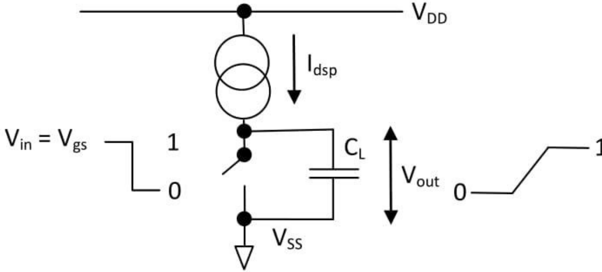
$$T_d = \left(1 + \frac{Z_{pu}}{Z_{pd}}\right)\tau$$

Thus, the inverter pair delay for inverters having 4:1 ratio is 5τ (which should be multiplied by a suitable factor to allow for wiring).

## ➢ Formal Estimation of CMOS inverter delay:

A CMOS inverter in general either charges or discharges a capacitive load $C_L$ and rise time $\tau_r$, or fall time $\tau_f$ can be estimated from the following analaysis:

1. Rise time estimation:
   Here, we assume that the p- device stays in saturation for the entire charging period of the load capacitor $C_L$. The circuit may then be modelled as shown.




The saturation current for the p- transistor is given as,

$$I_{dsp} = \frac{\beta_p (V_{gs} - |V_{tp}|)^2}{2}$$

This current charges $C_L$ and, since its magnitude is approximately constant, we have

$$V_{out} = \frac{I_{dsp}\, t}{C_L}$$

Substituting $I_{dsp} = \frac{\beta_p (V_{gs} - |V_{tp}|)^2}{2}$ in $V_{out} = \frac{I_{dsp}\, t}{C_L}$, we get

$$V_{out} = \frac{\frac{\beta_p (V_{gs} - |V_{tp}|)^2}{2} t}{C_L}$$

$$t = \frac{2\, C_L V_{out}}{\beta_p (V_{gs} - |V_{tp}|)^2}$$

Assuming that $t = \tau_r$ when $V_{out} = +V_{DD}$, so that

$$\tau_r = \frac{2\, C_L V_{DD}}{\beta_p (V_{DD} - |V_{tp}|)^2}$$

With $|V_{tp}| = 0.2\, V_{DD},\ then$

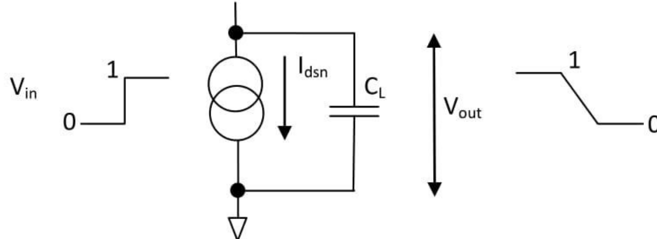
$$\boldsymbol{\tau_r = \frac{3\, C_L}{\beta_p\, V_{DD}}}$$

Algebraically,

$$\boldsymbol{\tau_r = 2.2\tau_p}$$

Therefore, the charging of $C_L$ is divided more correctly into two parts i.e. saturation and the resistive region of the transistor.

2. Fall- time estimation:




Similar reasoning can be applied for the discharge of $C_L$ through the p- transistor. Therefore, Similarly, we can write,

$$\tau_f = \frac{3\,C_L}{\beta_n\,V_{DD}}$$

Algebraically,

$$\tau_f = 2.2\tau_n$$

Therefore, we can summarize the inverter delay as:

$$\frac{\tau_r}{\tau_f} = \frac{\frac{3\,C_L}{\beta_p\,V_{DD}}}{\frac{3\,C_L}{\beta_n\,V_{DD}}} = \frac{\beta_n}{\beta_p}$$

3. Propagation Delay/ propagation time estimation:
   The propagation delay time $\tau_{propagation}$ is often used to estimate the 'reaction' delay time from input to output. When we use step- like input voltages, the propagation delay is defined by the simple average of two time- intervals.

$$\tau_{propagation} = 0.35\,(\tau_n + \tau_p)$$

**Factors which affect rise and fall times:**

1. $\tau_r$ and $\tau_f$ are proportional to $\frac{1}{V_{DD}}$.
2. $\tau_r$ and $\tau_f$ are proportional to $C_L$.
3. $\tau_r = \tau_f$ for equal n and p transistor geometries.

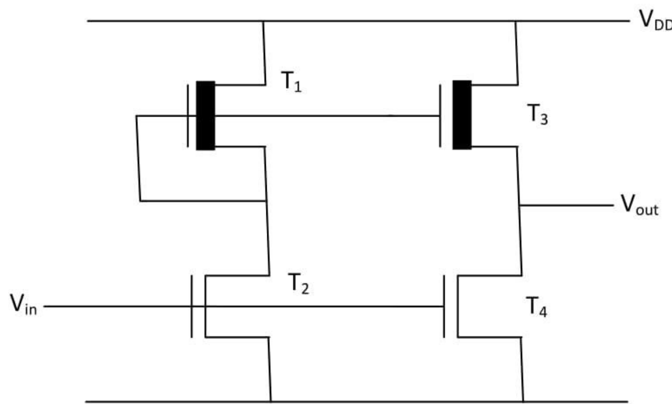
**Super buffer:**

A super buffer is a common alleviative approach for undesirable rise of delay problems of an conventional inverter/ inverter when it is used to drive more significant capacitive loads.

There are two types of super buffers:

1. Inverting type of super buffer
2. Non inverting type of super buffer

Inverting type super buffer (nMOS):

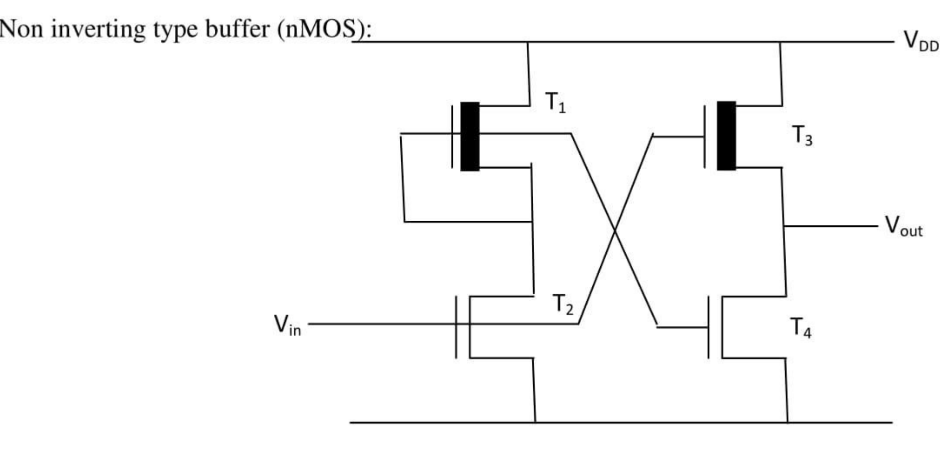



The inverting type as shown above is considered with a positive going logic transition $V_{in}$ at the input, it is seen that the inverter formed by $T_1$ and $T_2$ is turned ON and thus the gate $T_3$ is pulled down toward 0V with a small delay. Thus $T_3$ is cut off while $T_4$ (the gate of which is also connected to $V_{in}$) is turned ON and the output is pulled down quickly.

Now considering the opposite transition, when $V_{in}$ drops to 0V then the gate of $T_3$ is allowed to rise quickly to $V_{DD}$. Thus as $T_4$ is also turned OFF by $V_{in}$, $T_3$ is caused to conduct with $V_{DD}$ on its gate, that is, with twice the average voltage which would apply if the gate was tied to the source as in the conventional inverter.

Since $I_{ds}$ is directly proportional to $V_{gs}$, then it doubles the effective $V_{gs}$ will increase the current and thus reduce the delay in charging any capacitance on the output. Thus more symmetrical transitions are achieved.

Non inverting type buffer (nMOS):



The corresponding non inverting buffer as shown which has perspective structure of driving loads of 2 pF and with 5 nsec risetime.

If the inverting or non inverting buffer is arranged based on the native transistor, then it is known as native super buffer.

## Channel Length Modulation and Velocity Saturation:

The voltages exceeding the onset of saturation there is an effective decrease in the channel length of short channel transistor, this is referred as **channel length modulation**.

For example, the change in channel length $\Delta L$ for a n- transistor is approximated by,

$$\Delta L = \sqrt{\frac{2\ \varepsilon_0\ \varepsilon_{Si}}{q\ N_A}(V_{ds} - V_{th})}$$

And the resultant drain to source current $I^1{}_{ds}$ is approximated by,

$$I^1{}_{ds} = I_{ds}\frac{L}{L - \Delta L}$$

*Velocity Saturation:*

When the drain to source voltage of a short channel transistor exceeds a critical value, the charge carriers reach their maximum scattering limited velocity before pinch off. Thus less current is available from a short channel transistor than from a long channel transistor with similar width to length ratio and processing.

Therefore, channel length modulation and velocity saturation are the two effects important for short channel transistors, i.e. channel lengths $\leq 3\ \mu m$, and these effects should be taken into account.

## Fan-in and Fan- out:

The number of inputs to a logic gate in an inverter while adding complementary transistor pairs which increases the delay times as the capacitance of the transistor is increased is called **fan- in (FI)** and the number of gates is specified by the **fan- out (FO)** of the circuit. The fan- out gates acts as a load to the driving circuit because of their input capacitance.

**Problems:**

1. A resistor of value 100 kΩ needs to be made from a resistive layer of thickness 1µm. If the resistivity of the material is 1 Ωcm and the strip of width 5 µm is used, then what should be the length of the strip?

Sol.

Given:

$R = 100\ k\Omega = 1000\times10^3\ \Omega$

$\rho = 1\ \Omega cm = 1\times10^{-2}$

$t = 1\ \mu m = 1\times10^{-6}\ m$

$W = 5\ \mu m = 5\times10^{-6}\ m$

To find:

$L = ?$

WKT,

$$R = \frac{\rho\, L}{t\, W}$$

$$L = \frac{R\, t\, W}{\rho} = \frac{1000 \times 10^3 \times 1 \times 10^{-6} \times 5 \times 10^{-6}}{1 \times 10^{-2}} = 5 \times 10^{-5} m$$

Therefore, the length of the strip is $5 \times 10^{-5} m$ respectively.

2. A layer of MOS circuit has a resistivity of 1 Ωcm, a section of this material is 5 µm thick, 5 µm wide and has a length of 50 µm, calculate the resistance from one of the section to the other using the concept of sheet resistance.

Sol.

Given:

$\rho = 1\ \Omega cm = 1\times10^{-2}$

$t = 5\ \mu m = 5\times10^{-6}\ m$

$W = 5\ \mu m = 5\times10^{-6}\ m$

$L = 50\ \mu m = 50\times10^{-6}\ m$

To find:

R = ? using $R_s$, so first finding $R_s$ also

WKT,

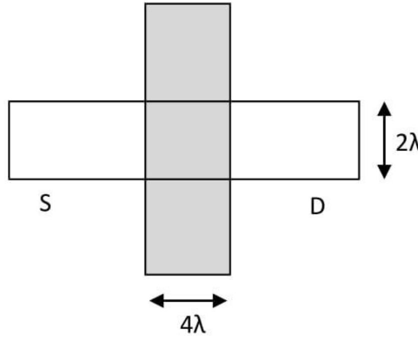
$$R_s = \frac{\rho}{t} = \frac{1 \times 10^{-2}}{5 \times 10^{-6}} = 0.2 \times 10^4 \ ohm/_{square}$$

And

$$R = R_s \frac{L}{W} = 0.2 \times 10^4 \times \frac{50 \times 10^{-6}}{5 \times 10^{-6}} = 2 \times 10^4 \Omega$$

Therefore, the value of resistance is $2 \times 10^4\ \Omega$ respectively.

3. For the given transistor structure, calculate the channel resistance in 5 µm, 2 µm and 1.2 µm technologies?




Sol.

Given:

L = 4λ

W = 2λ

- For nMOS:
  - In 5 µm technology

  WKT,

$$R = R_s \times \frac{L}{W} = 1 \times 10^4 \times \frac{4\lambda}{2\lambda} = 2 \times 10^4 \Omega$$

- In 2 µm technology

  WKT,

$$R = R_s \times \frac{L}{W} = 2 \times 10^4 \times \frac{4\lambda}{2\lambda} = 40\ k\Omega$$

- In 1.2 µm technology

  WKT,

$$R = R_s \times \frac{L}{W} = 2 \times 10^4 \times \frac{4\lambda}{2\lambda} = 40\ k\Omega$$

➢ For pMOS:

- In 5 µm technology

  WKT,

$$R = R_s \times \frac{L}{W} = 2.5 \times 10^4 \times \frac{4\lambda}{2\lambda} = 50\ k\Omega$$

- In 2 µm technology

  WKT,

$$R = R_s \times \frac{L}{W} = 4.5 \times 10^4 \times \frac{4\lambda}{2\lambda} = 90\ k\Omega$$

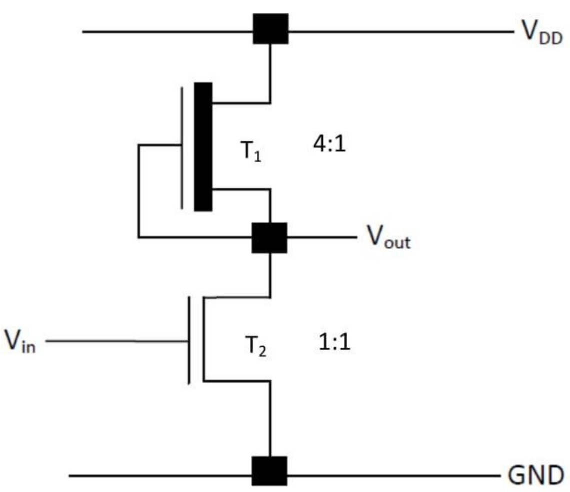
- In 1.2 µm technology

  WKT,

$$R = R_s \times \frac{L}{W} = 4.5 \times 10^4 \times \frac{4\lambda}{2\lambda} = 90\ k\Omega$$

Therefore, the channel resistance of the given transistor are found.

4. For the given nMOS inverter, calculate the total resistance in 5 µm and 2 µm techmologies.




Sol.

Given:

The inverter has two transistors $T_1$ with L = 4 and W = 1 and transistor $T_2$ with L = 1 and W = 1.

- In 5 µm technology

  WKT,

$$R = R_s \times \frac{L}{W}$$

$$R_{Total} = R_{T_1} + R_{T_2}$$

$$R_{Total} = \left(1 \times 10^4 \times \frac{4}{1}\right) + \left(1 \times 10^4 \times \frac{1}{1}\right) = 50\ k\Omega$$

- In 2 µm technology

  WKT,

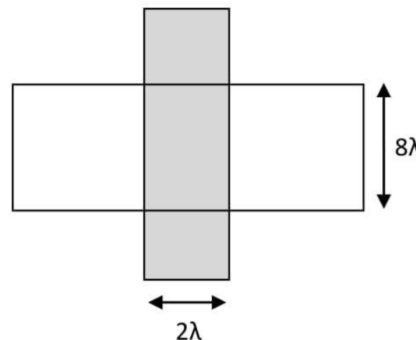
$$R = R_s \times \frac{L}{W}$$

$$R_{Total} = R_{T_1} + R_{T_2}$$

$$R_{Total} = \left(2 \times 10^4 \times \frac{4}{1}\right) + \left(2 \times 10^4 \times \frac{1}{1}\right) = 100\ k\Omega$$

Therefore, the total resistance of the inverter in 5 µm technology is 50 kΩ and in 2µm technology is 100 kΩ respectively.

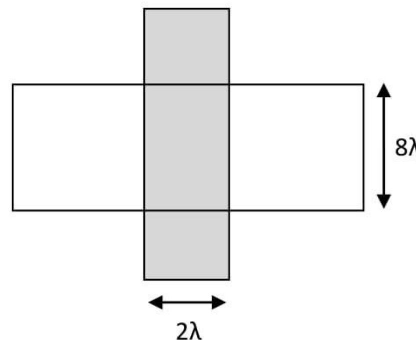
(b) For the given transistor structure, calculate the channel resistance in 5 µm, 2 µm and 1.2 µm technologies?




(c) Calculate the total resistance in a CMOS inverter in 5 µm, 2 µm and 1.2 µm technologies?

(Note/ Hint : For CMOS inverter $L:W = 1:1$ $(i.e.)$ $\frac{L}{W} = \frac{1}{1}$)

(b) For the given transistor structure, calculate the channel resistance in 5 µm, 2 µm and 1.2 µm technologies?




(c) Calculate the total resistance in a CMOS inverter in 5 µm, 2 µm and 1.2 µm technologies?

(Note/ Hint : For CMOS inverter $L:W = 1:1$ $(i.e.)$ $\frac{L}{W} = \frac{1}{1}$)

# UNIT-3

# SUBSYSTEM DESIGN

## 1. COMBINATIONAL LOGIC:

The circuits in which the output/ outputs depends on the present input are referred as combinational logic circuits. Examples of such circuits are adders, subtractors, multipliers, shifters etc. In dense integration the major arithmetic logic circuits are adders, multipliers and shifters.

### ADDERS:

An adder is an input- output logical circuit which performs addition operation. Addition is the most commonly used arithmetic operation, often is the speed- limiting element as well. Therefore careful optimization of the adder is important; this optimization can proceed at the logic (using Boolean equations) or circuit level (using transistor sizing and circuit topology).

Some optimized adder circuits are:

1. Carry Select Adder
2. Carry Bypass Adder
3. Carry Look Ahead Adder
4. Manchester Carry Chain Adder etc

### The Binary Adder:

It is a three input (*A, B and* $C_{in}$), two output (*Sum S and Carry Out* $C_0$) combination logic circuit which performs the arithmetic addition operation, usually referred as the binary adder or full adder.

$$S = A \oplus B \oplus C_i$$
$$= A\bar{B}\bar{C}_i + \bar{A}B\bar{C}_i + \bar{A}\bar{B}C_i + ABC_i$$
$$C_o = AB + BC_i + AC_i$$

| 0 | 1 | 0 | 1 | 0 | propagate |
|---|---|---|---|---|---|
| 0 | 1 | 1 | 0 | 1 | propagate |
| 1 | 0 | 0 | 1 | 0 | propagate |
| 1 | 0 | 1 | 0 | 1 | propagate |
| 1 | 1 | 0 | 0 | 1 | generate |
| 1 | 1 | 1 | 1 | 1 | generate |

The sum and carry can be expressed as *Propagate (P), Generate (G) and Delete (D)* for all the input binary logical combinations depicted by the carry status as shown above. The three variables *P, G and D* depend on the inputs A and B only.

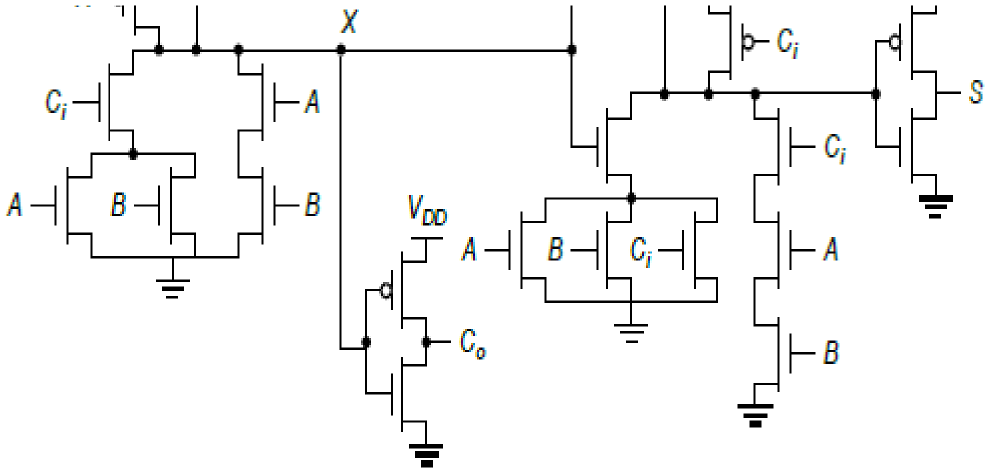
(Sometimes P = A+B is also used)

Expressing $C_0$ and S using *P* and *G*,

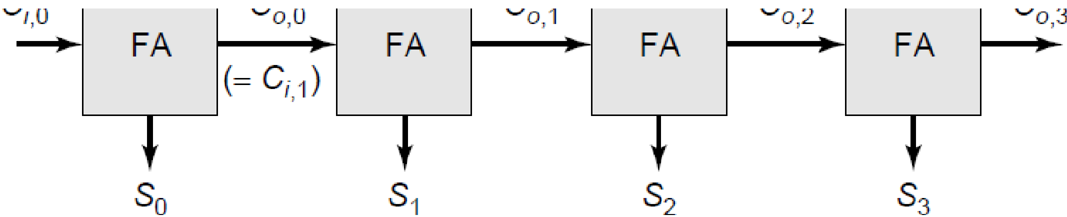
$$C_o(G,P) = G + PC_i$$

$$S(G,P) = P \oplus C_i$$

Expressions of $C_0$ and S can also be derived using *D* and *P* respectively.




***COMPLIMENTARY STATIC CMOS FULL ADDER USING 28 TRANSISTORS***




***FOUR BIT RIPPLE CARRY ADDER TOPOLOGY***

An N bit adder can be constructed by cascading N full adder (FA) circuits in series. Connecting $C_{o,\ k-1}$ to $C_{i,\ k}$ for $k = 1$ to $N-1$, and the first carry- in $C_{i,0}$ to 0 as shown in figure. This configuration is called **ripple- carry adder**, since the carry bit ripples from one stage to the other. The delay through the circuit depends upon the number of logic stages that must be traversed and is a function of the applied input signals.

For some input signals no rippling effect occurs at all. While for others the carry has to ripple all the way from LSB to the MSB. The propagation delay of such structure also called as critical path is defined as the worst case delay over all possible input paterns.

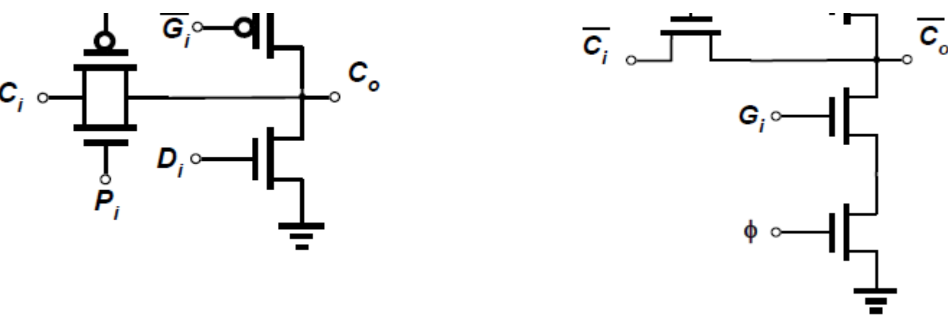
In ripple carry adder the worst case delay happens when a carry generated at least significant bit (LSB) position propagates all the way to the most significant bit (MSB) position. This carry is finally consumed in the last stage to produce the sum. The delay is then proportional to the number of bits in the input words N and is approximated by,

$$t_{adder} \approx (N-1)t_{carry} + t_{sum}$$

*Where: $t_{carry}$ and $t_{sum}$ equal the propagation delays from $C_i$ to $C_o$ and $S$*

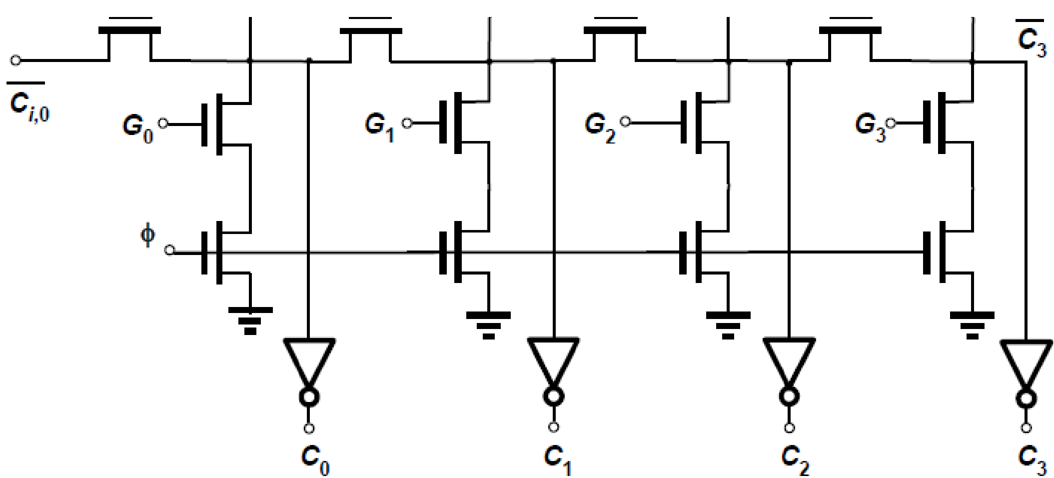
## Manchester Carry Chain Adder:

The Manchester carry chain adder uses cascading of pass transistors to implement the carry chain. It can be implemented operating as a static or dynamic. In static implementation it uses propagate, generate and kill/ delete signals and in dynamic implementation it uses only propagate and generate signals.




STATIC LOGIC IMPLEMENTATION

DYNAMIC LOGIC IMPLEMENTATION

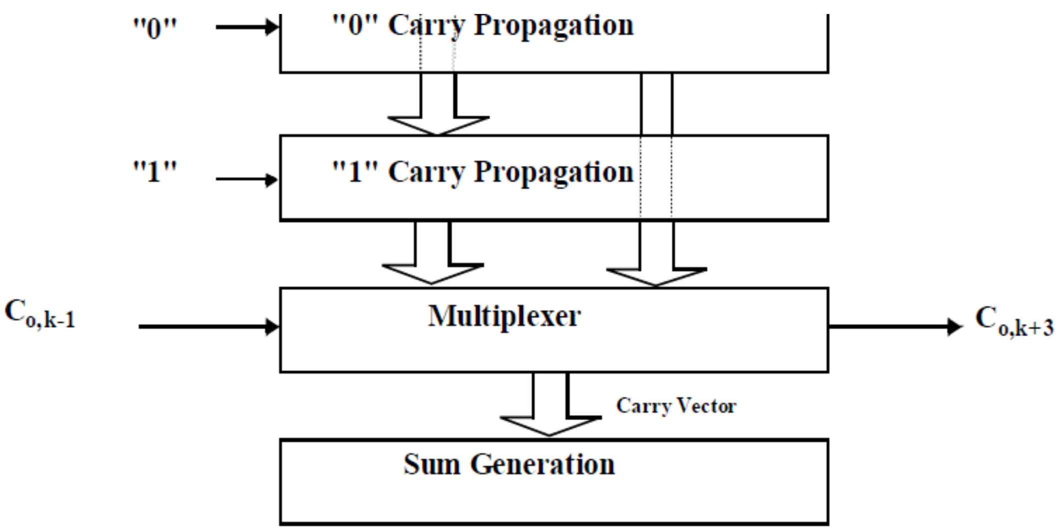



MANCHESTER CARRY CHAIN ADDER (IN DYNAMIC LOGIC) (FOUR- BIT)

The Manchester carry chain adder can be seen in the above figure in dynamic logic. During the pre charge phase ($\Phi = 0$), all the intermediate nodes of the pass transistor carry chain are pre charged to $V_{DD}$. During evaluation, the $A_k$ node is discharged when there is a incoming carry and the propagate signal $P_k$ is high, or when the generate signal for stage k ($G_k$) is high.

**Carry Select Adder:**

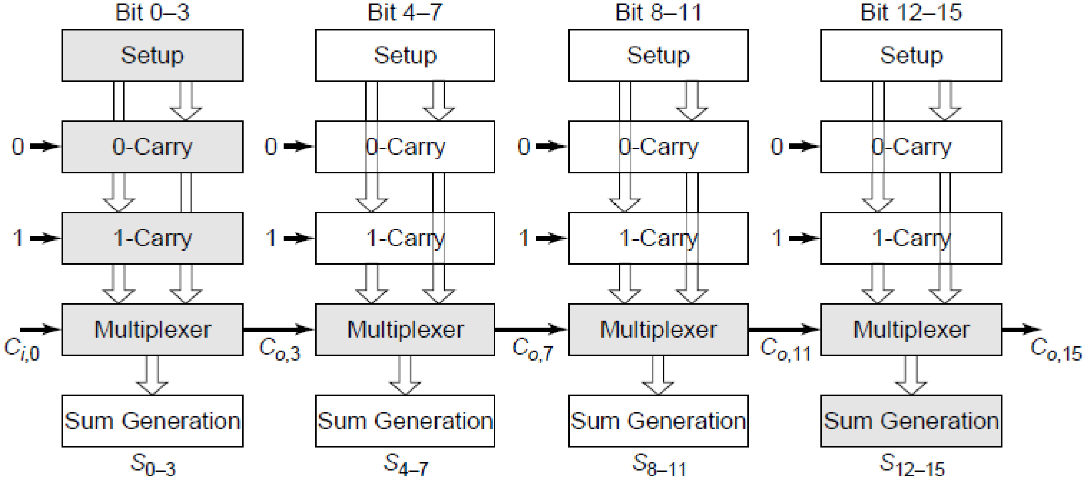
Carry select adder overcomes the problem of delay/ waiting for a carry in a ripple carry adder, where in the ripple carry adder- each adder cell has to wait for the incoming carry before an outgoing carry can be generated. This is done by anticipating both possible values of the carry input and evaluate the result for both possibilities in advance. Once the real value of the incoming carry is known, the correct result is easily selected with a simple multiplexer stage. This implementation is referred as **Carry- Select Adder**.




*FOUR BIT CARRY SELECT ADDER MODULE TOPOLOGY*

As seen from the above figure which illustrates a four bit Carry Select Adder. Consider the block of adders, which is adding bits $k$ to $k+3$, instead of waiting on the arrival of the output carry of the bit $k-1$, both the 0 and 1 possibilities are analyzed. From a circuit point of view this means that two carry paths are implemented. When $C_{0,\,k-1}$ finally settles, either the result of the 0 or 1 path is selected by the multiplexer, which can be performed with a minimal delay. As it is evident from the figure, the hardware overhead of the carry select adder is restricted to an additional carry path and a multiplexer, and equals about 30% with respect to a ripple carry structure.

A full carry select adder can be constructed by chaining a number of equal length adder stages, as in the carry- bypass approach.




***SIXTEEN BIT CARRY SELECT ADDER (CRITICAL PATH IS SHOWN BY SHADING)***

From the inspection of the circuit we can derive a first order model of the worst case propagation delay of the module as,

$$t_{add} = t_{setup} + Mt_{carry} + \left(\frac{N}{M}\right)t_{mux} + t_{sum}$$

*Where:*

*$t_{setup}$, $t_{sum}$ and $t_{mux}$ are fixed delays*

*N and M are the total number of bits and the number of bits per stage*

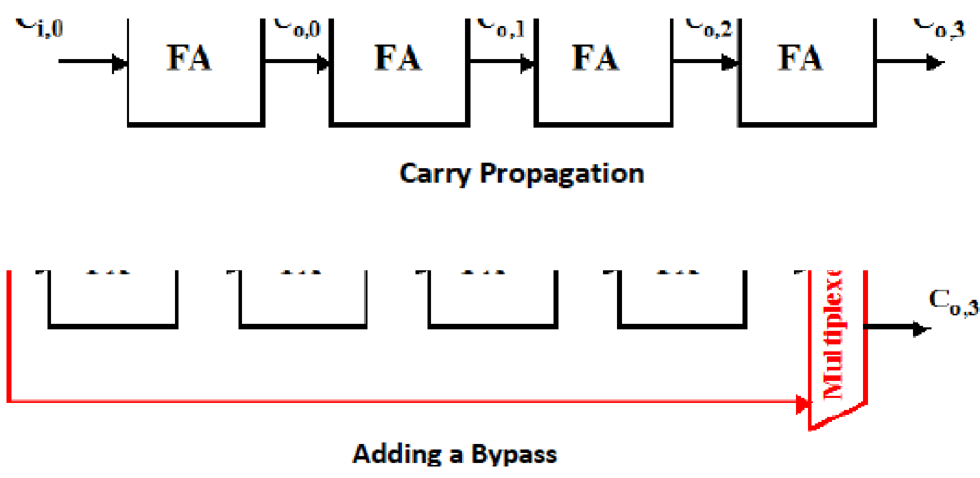
*$t_{carry}$ is the delay of the carry through a single full adder cell*

The carry delay through a single block is proportional to the length of that stage or equals $M\ t_{carry}$.

The propagation delay of the adder is again linearly proportional to *N*. The reason for this linear behaviour is that the block- select signal that selects between the 0 and 1 solution still has to ripple through all stages in the worst case.

## Carry Skip Adder:

A carry-skip adder (also known as a carry-bypass adder) is an adder implementation that improves on the delay of a ripple-carry adder with little effort compared to other adders. The improvement of the worst-case delay is achieved by using several carry-skip adders to form a block-carry-skip adder.

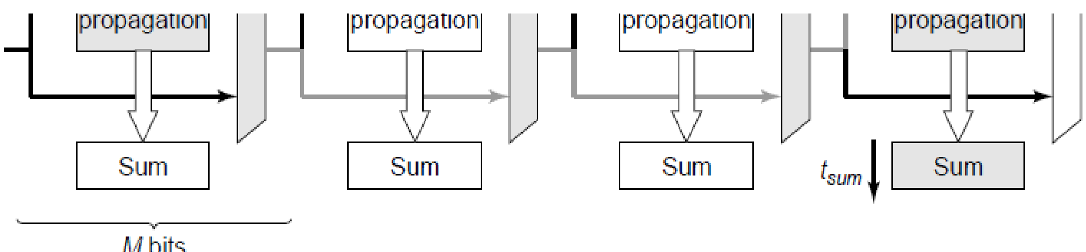



*BASIC CARRY BYPASS/ SKIP ADDER STRUCTURE*

Consider the four bit adder block as shown in figure "Carry Propagation". Let the values of $A_k$ and $B_k$ $(k=0...3)$ are such that all propagate signals $P_k$ $(k=0...3)$ are high. An incoming carry $C_{i,0}=1$ propagates under those conditions through the complete adder chain and causes an outgoing carry $C_{o,3}=1$. i.e.

$$If\ (P_0P_1P_2P_3 = 1)\ then\ C_{o,3} = C_{i,3}\ else\ either\ \textbf{Delete}\ or\ \textbf{Generate}\ occured$$

This can be used to speed up the operation of the adder as shown in figure "Adding a Bypass". When $BP = P_0\ P_1\ P_2\ P_3 = 1$, the incoming carry is forwarded immediately to the next block through the bypass transistor $M_b$, hence the name carry- bypass or carry- skip adder.




*SIXTEEN BIT CARRY BYPASS ADDER (WORST CASE SHOWN SHADED)*

The delay of a 16 bit carry bypass adder is done by assuming that the total adder is divided in (N/M) equal length bypass stages, each of which contains M bits. An approximate expression for the total propagation time can be derived from the above shown figure as,

$$t_p = t_{setup} + Mt_{carry} + \left(\frac{\ }{M} - 1\right) t_{bypass} + (M-1)t_{carry} + t_{sum}$$

*Where:*

$t_{setup}$ *is the fixed overhead time to create the generate and propagate signals*

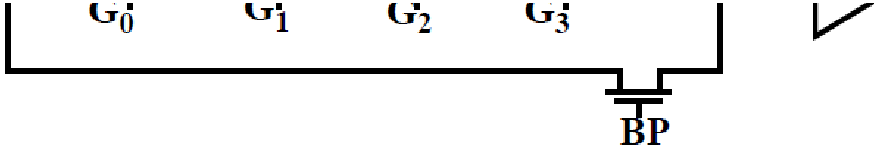
$t_{carry}$ *is the propagation delay through a single bit. The worst case carry-propagation delay through a single stage of M bits is approximately M times larger.*

$t_{bypass}$ *is the propagation delay through the bypass*

*multiplexer of a single stage* $t_{sum}$ *is the time to*

*generate the sum of the final stage*

**Carry Bypass in Manchester Carry- Chain Adder:**




MANCHESTER CARRY CHAIN IMPLEMENTATION OF BYPASS ADDER

As seen from the figure, it shows the possible carry- propagation paths when the full adder circuit is implemented in Manchester Carry style. This figure shows how the bypass speeds up the addition. The carry propagates either through the bypass path or a carry is generated somewhere in the chain. In both cases the delay is smaller than the normal ripple configuration. The area overhead incurred by adding the bypass path is small and typically ranges between 10 and 20%, however adding the bypass path breaks the regular bit- slice structure.

**<u>MULTIPLIERS</u>:**

A multiplier is an in effect complex adder arrays which performs multiplication operation. These multiplication operations are expensive and slow. The performance of many computational problems often is dominated by the speed at which a multiplication operation can be executed. This has promoted the integration of

complete multiplication units in state of the art digital signal processors and microprocessors.

The analysis of the multiplier gives us an further insight into how to optimize the performance (or the area) of complex circuit topologies.

Consider two unsigned binary numbers A and B that are M and N bits wide respectively. To introduce the multiplication operation it is useful to express A and B in the binary form.

$$A = \sum_{i=0}^{M-1} A_i 2^i \ and \ B = \sum_{j=0}^{N-1} B_j 2^j$$

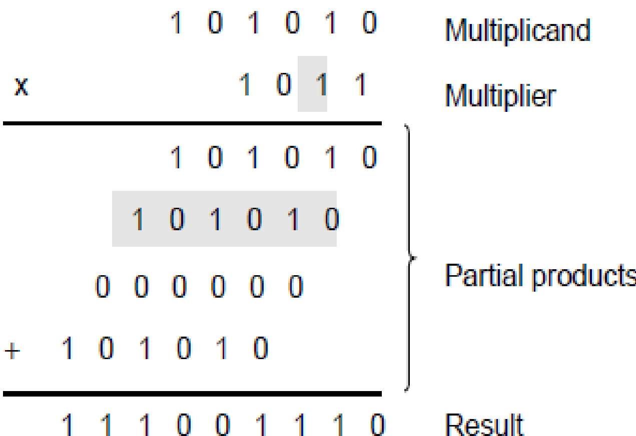
With $A_i$ and $B_j \in \{0,1\}$. The multiplication operation is then defined as:

$$C = A \times B = \sum_{k=0}^{M+N-1} Z_k 2^k$$

$$= \left(\sum_{i=0}^{M-1} A_i 2^i\right)\left(\sum_{j=0}^{N-1} B_j 2^j\right) = \sum_{i=0}^{M-1}\left(\sum_{j=0}^{N-1} A_i B_j 2^{\,i+j}\right)$$

The simplest way to perform a multiplication is to use a single two input adder, for inputs M and N bit wide the multiplication takes M cycles using N bit adder. This *shift and add algorithm* for multiplication adds together M partial products. Each partial product is generated by multiplying the multiplicand with a bit of the multiplier which essentially is an AND operation, and by shifting the result on the basis of the multipliers bit's position.

A faster way of implementing multiplication is to resort to an approach similar to manually computing a multiplication. All the partial products are generated at the same time and organized in an array. A multioperand addition is applied to compute the final product as shown below.



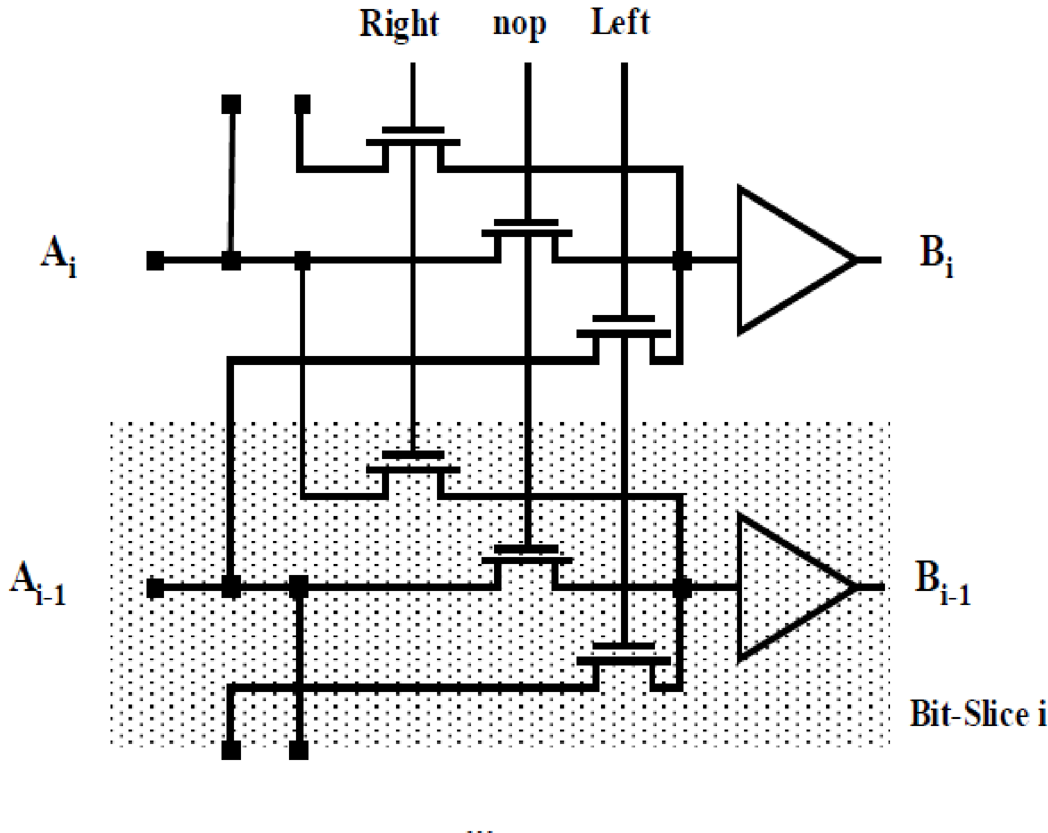

***AN EXAMPLE OF BINARY MULTIPLICATION***

As shown above. This set of operations can be mapped directly into hardware. The resulting structure is called an ***array multiplier*** and combines the *partial- product generation, partial- product accumulation and final addition* functions.

## SHIFTERS:

The shifter performs the shifting operation which is an essential arithmetic operation that requires adequate hardware. A shifter shifts a data word left or right over a constant amount and is implemented by an appropriate signal wiring. Latter can be implemented as a combination of add and shift operations. Shifters are extensively used in floating point units, scalars and multiplications by constant numbers.

Programmable shifters are more complex and require active circuitry, these are nothing less than an intricate multiplexer circuit.



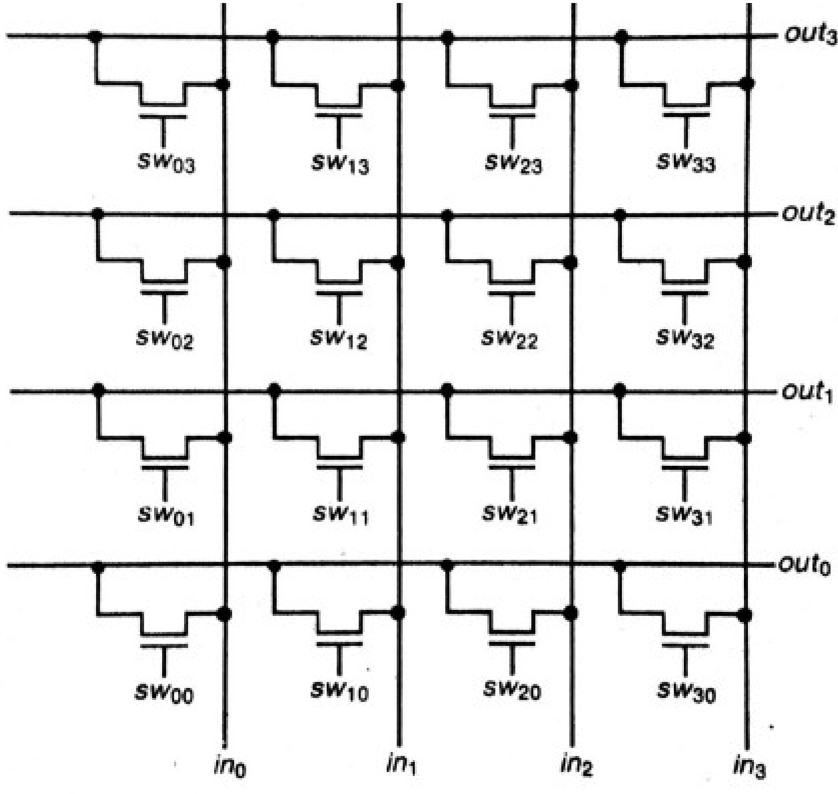

***ONE BIT PROGRAMMABLE SHIFTER (LEFT- RIGHT SHIFTING)***

The figure shows a simple one bit shifter. Depending on the control signals, the input word line is either shifted left or right, or else remains unchanged. Multi bit shifters can be built by cascading a number of these units. Multi bit shifters are complex and slow for larger shift values.

There are two types of shifters:

- Crossbar Shifter
- Barrel Shifter

**<u>Crossbar Shifter</u>:**

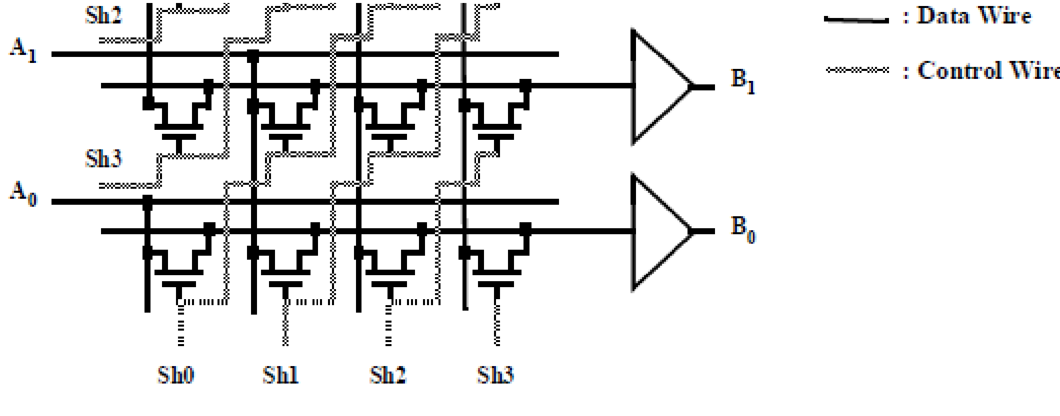



A 4 × 4 CROSSBAR SHIFTER

The structure of the four bit shifter known as crossbar shifter is shown, which consists of 16 transistors connected in lattice like structure with 16 control signals, one for each transistor must be provided to drive the crossbar shifter. To carry out the shift operation four control signals have to be activated, then the corresponding transistor switches is turned ON and the data gets shifted to right by one place. This shifter requires 16 control signals as the number of bits are increased, control signals are also increased. If the control signal due to fault or error gets missed, then all the transistor switches might get turn ON, all inputs connected to all outputs, resulting in short circuit.

Therefore to avoid the problem happening in crossbar shifter an adaptation of this arrangement that we can couple the switch gates together in groups of four as its a four bit shifter and also form four separate groups corresponding to shift of zero, one, two and three bits. This arrangement is readily adapted so that the in- lines also run horizontally, this arrangement is known as *Barrel Shifting or the shifter is known as Barrel Shifter*.

**Barrel Shifter:**




*A 4 × 4 BARREL SHIFTER*

The structure of the barrel shifter is shown in figure. It consists of an array of transistors, in which the number of rows equals the word length of the data, and the number of columns equals the maximum shift width. Here in the structured considered both are equal set to four. The control wires are routed diagonally through the array. The barrel shifter is an overcome of the drawbacks of a crossbar shifter.

The main advantage of this shifter is that the signal has to pass through at most one transmission gate, in other words the propagation delay is theoretically constant and independent of the shift value or shifter size. This is not practical as the capacitance at the input of the buffers rises linearly with the maximum shift width.

Here the layout size of the circuit is not dominated by the active transistors as in the case of all other arithmetic circuits but by the number of wires running through the cell. The size of the cell is bounded by the pitch of the metal wires. Barrel shifter needs control signals to shift, and the number of control signals depends on the number of bits i.e. here for a four bit barrel shifter, four control signals are required and the barrel shifter requires an extra decoder to decode the signal while shifting into the former when required.

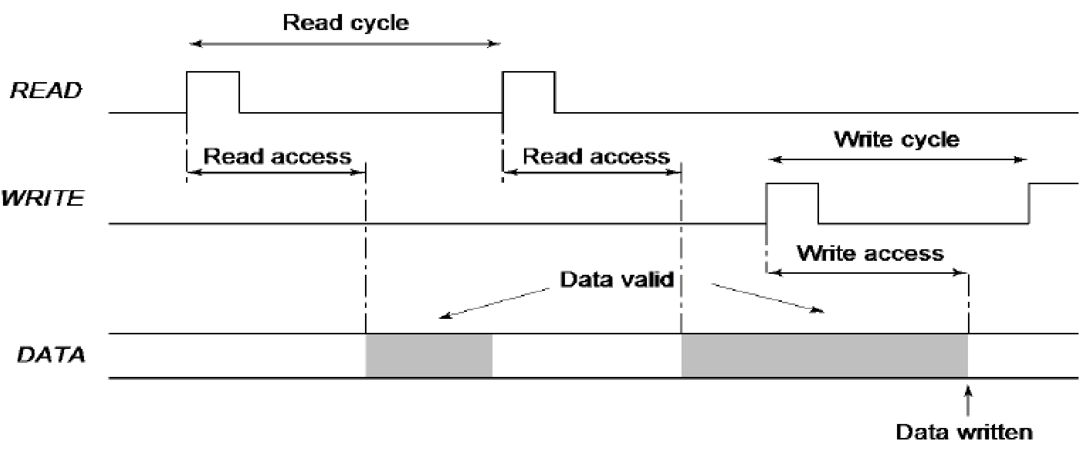
## High Density Memory Elements

### *CLASSIFICATION OF MEMORY*:

Memories are used for storing large data in digital system. The memory requirement is dependent on the application. Basically memory is made of transistor cells, for large storage dense number of transistor integration is done and to reduce the number of transistors different MOS technologies are been used. Memory is classified into three categories:

1. Read- Write Memory (RWM)
2. Non- volatile Read Write Memory (NVRWM)
3. Read- Only Memory (ROM)

| RWM | | NVRWM | ROM |
|---|---|---|---|
| **Random Access** | **Non- Random Access** | EPROM<br>$E^2$PROM<br>FLASH | Mask- programmed<br>programmable (PROM) |
| SRAM<br>DRAM | FIFO<br>LIFO<br>Shift register<br>CAM | | |




*MEMORY TIMING*

## READ- WRITE MEMORY:

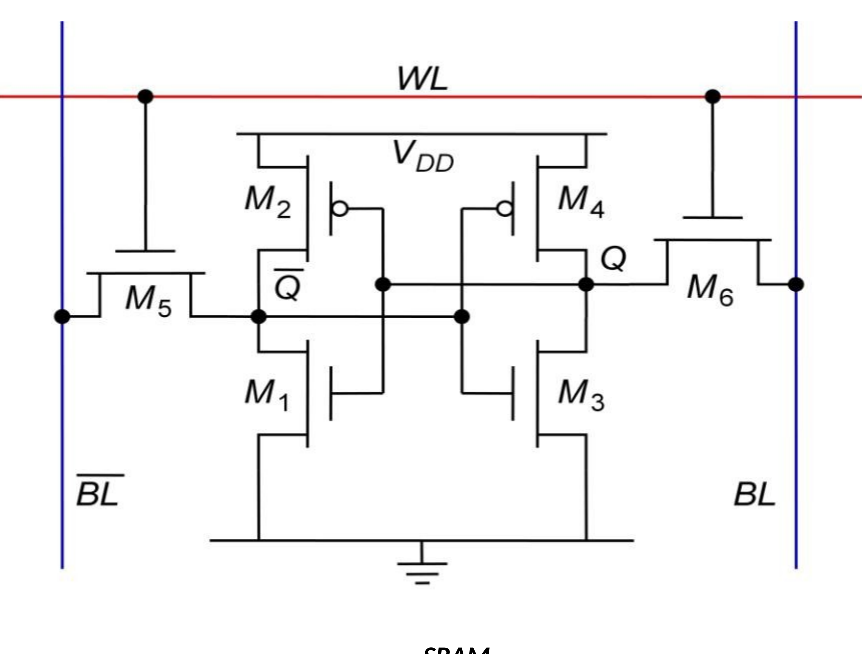
As prerequisite from the classification of memory RWM comprises of **Static Random Access Memory (SRAM)** and **Dynamic Random Access Memory (DRAM)**.

### Static Random Access Memory (SRAM) (also called as 6T SRAM):

In SRAM the data is stored as long as the supply is applied. It is fast and differential.

The SRAM consists of 6 transistors (6T) i.e. a combination of two CMOS inverters comprising 4T ($M_1$, $M_2$, $M_3$ and $M_4$) and two nMOS pass transistors ($M_5$ and $M_6$), totally constituting 6T.

In SRAM access to each cell is enabled by the word line, which replaces the clock and controls the two pass transistors $M_5$ and $M_6$, shared between the read and write operation.




SRAM

Operation:

The SRAM should be sized as small as possible to achieve high memory densities.

*Performing Write Operation:*

To write the data into the cell, again the word line is enabled, the data to be written is made available in the bit lines and the data is stored in the latch.

*Performing Read Operation:*

To read the data from the cell, word line is enabled, this makes transistors $M_5$ and $M_6$ ON. Hence the stored data is available in both the true and complemented form in the bit line and complement bit line respectively.

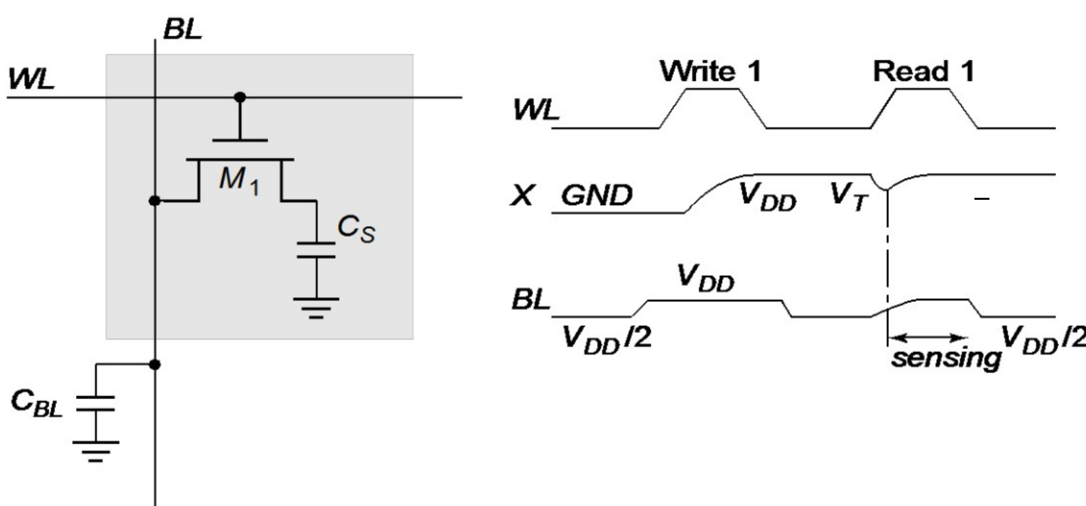
**Dynamic Random Access Memory (DRAM):**

In DRAM periodic refreshment is required, these are small consisting of IT, 3T or 4T cells. DRAMs are slower and single ended. In DRAM the data is stored in the parasitic capacitance which discharges with time.

The various types of DRAM are:

1. 1Transistor/ 1T DRAM
2. 3T DRAM
3. 4T DRAM

**1T DRAM:**




*1T DRAM AND ITS SIGNAL WAVEFORMS DURING READ AND WRITE OPERATION*

The 1T DRAM consists of one transistor ($M_1$) which is connected to the word line (WL) and bit line (BL) along with a storage capacitor.

Operation:

*Performing Write Operation:*

While performing the write operation the word line is enabled and the data from the bit line is stored in the capacitor $C_S$.

*Performing Read Operation:*

While performing the read operation the word line is enabled and stored data becomes available at the bit line. The charge distribution takes place between the bit line and storage capacitor.

- The 1T DRAM is mostly used in high density DRAM architecture
- The voltage swing is small, typically around 250 mV and it requires a sense amplifier for each bit line, due to charge distribution read out.

## 3T DRAM:

*3T DRAM AND ITS SIGNAL WAVEFORMS DURING READ AND WRITE OPERATION*

The three transistor DRAM consists of three transistors $M_1$, $M_2$ and $M_3$ respectively as shown. Transistor $M_1$ is connected to the write word line (WWL), the transistor $M_3$ is connected to the Read Word line (RWL) and the transistors $M_2$ is used for storing the binary data in association with its storage capacitance $C_S$.

Operation:

*Performing Write Operation:*

While performing the write operation, the writing word line (WWL) is enabled, the logic from the write bit line is passed to the parasitic storage capacitance $C_S$.

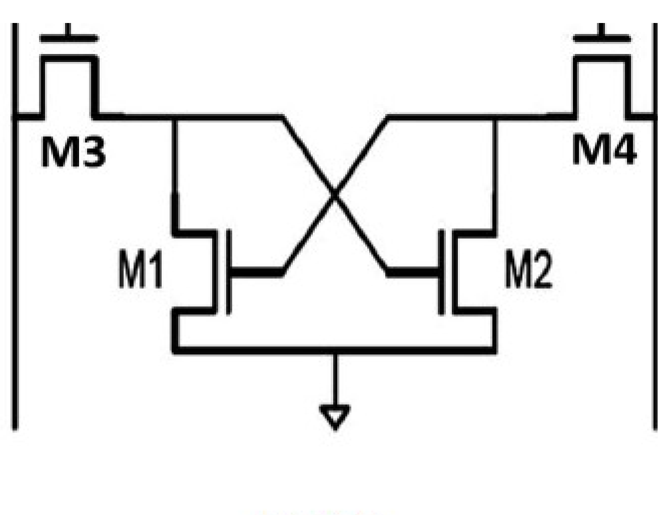
*Performing Read Operation:*

While performing the read operation, the read word line in enabled (RWL) and the complement of the stored data becomes available in the read bit line.

- In contrast to SRAM, in 3T DRAM no constraints exist on the device ratios.
- In contrast to other DRAM cells, reading the 3T cell contents is non-destructive i.e. the data value stored in the cell is not affected by a read.
- Here no special process steps are needed. The storage capacitance is nothing more than the gate capacitance of the readout device, this is in contrast to other DRAM cells.
- The 3T DRAM is attractively used in embedded memory applications.
- The value stored on the storage node X when writing a 1 equals to $V_{WWL} – V_{th}$. This threshold loss reduces the current flowing through $M_2$ during a read operation and increases the read access time.

- To prevent this, some designs bootstrap the word line voltage i.e. raise $V_{WWL}$ to a value higher than $V_{DD}$

**4T DRAM:**




*4T DRAM*

The 4T DRAM consists of four transistors $M_1$, $M_2$, $M_3$ and $M_4$ respectively with connections to word and bit lines as shown.

Here the binary data is stored in the parasitic capacitances present between the transistors $M_1$, $M_3$ and $M_2$, $M_4$ with respect to ground. The data is written in complemented form by enabling the word line. As there is no restoring path from $V_{DD}$ to these capacitances, the stored charge is lost. So to retain the logic level, the capacitors must be refreshed periodically.

During read operation, the word line (WL) is enabled and the stored data becomes available at the bit lines (BLs) both in the true and complemented form respectively.

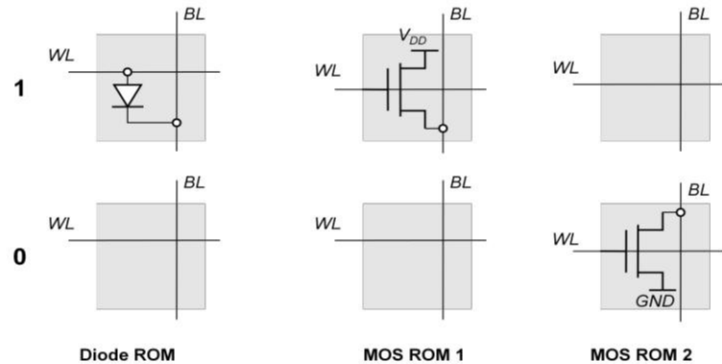
## READ ONLY MEMORY (ROM):

Using ROM several large numbers of potential applications can be accomplished. Programs for processors with fixed applications such as washing machines, calculators and gaming machines- once developed and debugged need only reading.

Fixing the contents at manufacturing time leads to small and fast implementations.

With the fact that ROM cells are permanently fixed considerably simplifies the design. The cell should be so that a 0 or 1 is present to the bit line upon activation of its word line.

The different approaches for implementing 1 and 0 ROM cells are shown below.

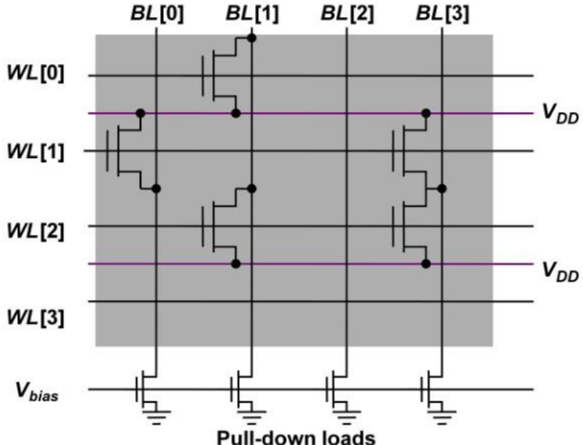



*DIFFERENT APPROACHES FOR IMPLEMENTING 1 AND 0 ROM CELLS*

Using the ROM cells, ROM arrays can be developed which can be designed as:

1. OR based ROM Memory Design
2. NOR based ROM Memory Design
3. NAND based ROM Memory Design

## OR BASED ROM MEMORY DESIGN:

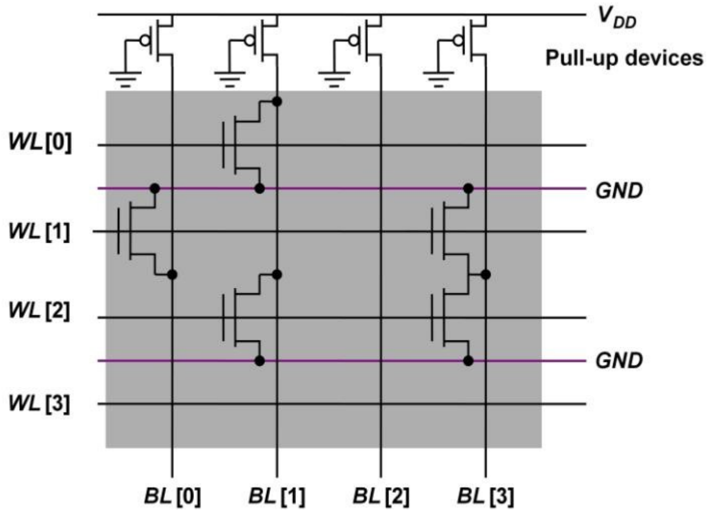



*A 4 × 4 OR ROM CELL ARRAY*

The figure shows a 4 × 4 OR RAM cell array. Here the values of the data stored at addresses 1, 2 and 3 as shown in figure are determined.

Here the absence of transistors between the word lines and bit lines means that logic 1 is stored and the 0 cell is realized by providing a MOS device between the bit lines and ground. Applying a high voltage on the word lines turns on the device, which in turn pulls down the bit line to ground.

**NOR BASED ROM MEMORY DESIGN:**




*A 4 × 4 NOR ROM CELL ARRAY*

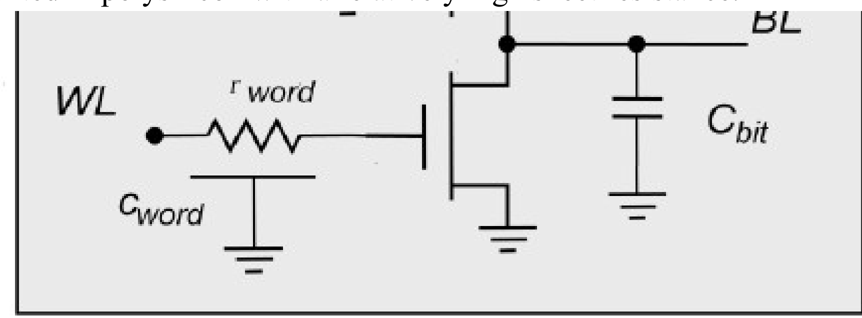
The above figure shows a 4 × 4 NOR ROM cell array. The values of the data stored at addresses 0, 1, 2 and 3 are determined. The combination of a bit line, PMOS pull- up and NMOS pull- downs constitutes a pseudo- NMOS NOR gate with the word line as inputs. Therefore, an N × M ROM memory can be considered as a combination of M NOR gates with at most N inputs (for a fully populated column) are called a NOR ROM.

Under normal operating conditions only one of the word line goes high, and at most one of the pull- down devices is turned on. This raises the issues regarding the sizing of both the cell and pull- up transistors.

To keep the cell size and the bit line capacitance small, the pull- down device should be kept as close as possible to minimum size and the resistance of the pull- up device must be larger than the pull- down resistance to ensure an adequate low level.

**Equivalent model of NOR based memory design:**

The equivalent shown is appropriate for the analysis of the word and bit line delay of the NOR ROM. The word line is best modeled as a distributed RC line since it is implemented in polysilicon with a relatively high sheet resistance.

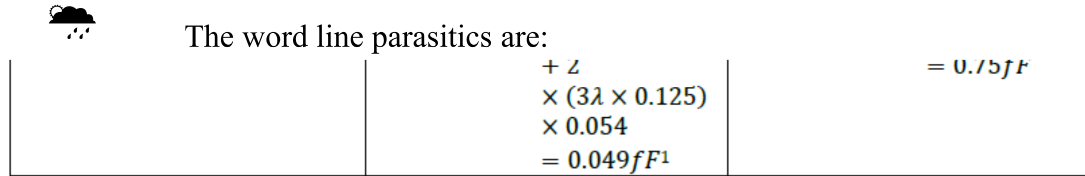



**EQUIVALENT MODEL FOR NOR BASED ROM**

The wire line parasitic consists of wire capacitance and gate capacitance, the wire resistance by the polysilicon and the bit line parasitic consists of resistance not dominant (metal) and drain, gate- drain capacitance.

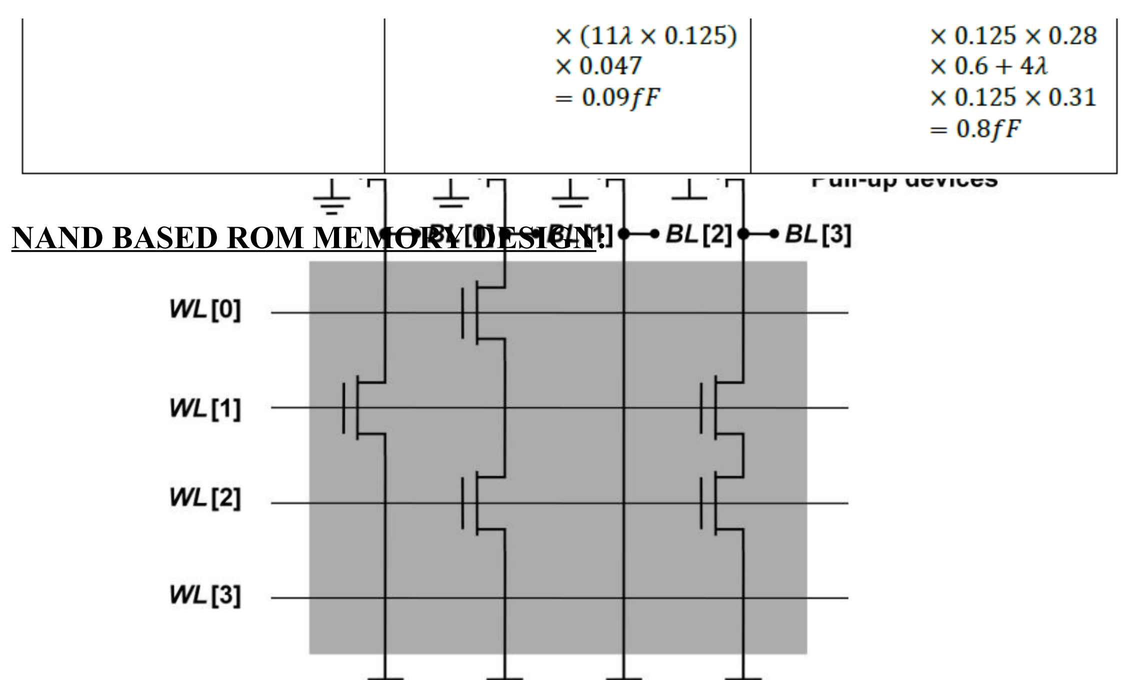
The bit line is implemented in Aluminum and the resistance of the line only comes into play for very long lines. It is reasonable to assume that here a purely capacitive model is adequate and that all capacitive loads connected to the wire can be lumped into a single element.

The word line parasitics are:



| | | |
|---|---|---|
| | $+ 2$ $\times (3\lambda \times 0.125)$ $\times 0.054$ $= 0.049 fF$ | $= 0.75 fF$ |

The bit line parasitics are:



| | | |
|---|---|---|
| | $\times (11\lambda \times 0.125)$ $\times 0.047$ $= 0.09 fF$ | $\times 0.125 \times 0.28$ $\times 0.6 + 4\lambda$ $\times 0.125 \times 0.31$ $= 0.8 fF$ |



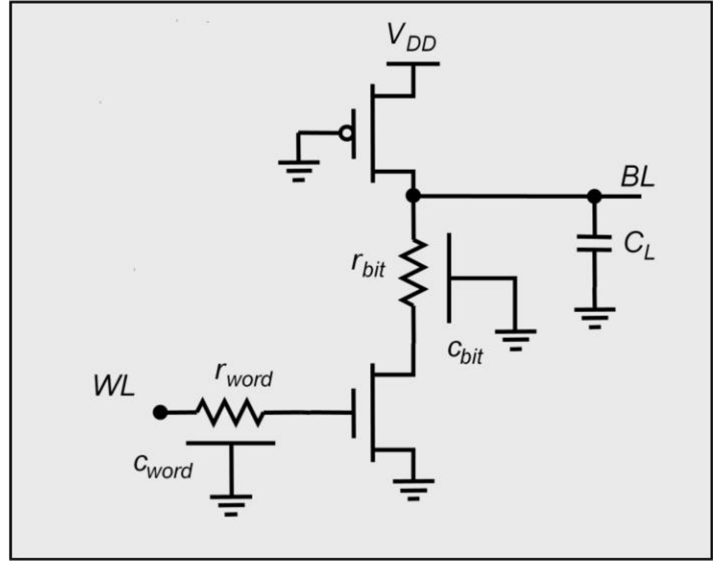
**NAND BASED ROM MEMORY DESIGN:**

***A4×4NANDROM***

The above fugue shows a 4 × 4 NAND ROM. The values of the data stored at addresses 0, 1, 2 and 3 are determined. The main advantage of using the NAND based structure is it that the basic cell consists of a transistor or no transistor and no connection to any of the supply voltages is needed, this reduces the cell size subsequently.

Here all word lines are high by default with the exception of selected row.

Equivalent neither model of NOR based memory design:




***EQUIVALENT MODEL OF NAND BASED ROM***

Using the above shown equivalent model for a NAND based ROM we can derive the delay of the NAND structure. It is identical to model the word line as the bit line behavior is complex due to its long chain of connected transistors. The worst behavior occurs when the transistor at the bottom of the chain is switched and the column is completely populated with transistors.

Each of the series transistors (which are normally in the ON mode) is modeled as a RC combination. The entire chain can be modeled as a distributed RC network for large memories.

The word line parasitic are:

➢ The bit line parasitics are:

| Resistance | Wire Capacitance | Gate Capacitance |
| --- | --- | --- |
| $= {}^{13}/_{4} \times 1.5 = 8.7k\Omega$ | $Included\ in\ diffusion\ capacitance$ | $= (3\lambda \times 3\lambda)(0.125)^2 \times 2 \times 0.56 + 2 \times 3\lambda \times 0.125 \times 0.28 \times 0.6 + (3\lambda \times 4\lambda)(0.125)^2 \times 6 = 0.85fF$ |

In contrast to the NOR based ROM, here in NAND based ROM the source/ drain capacitance must include the gate- source and gate- drain capacitances, which means the complete gate capacitance whereas in NOR based ROM only the drain-source overlap capacitance is included.

**Assignment:**

1. Explain Partial- product generation, partial- product accumulation and final addition in multipliers?
2. What is Pseudo- NMOS?
3. Differentiate between static and dynamic memories?
4. Draw the gate level logic diagram of a half adder, full adder and D- flip flop along with their truth tables.

## *ANALOG VLSI DESIGN:*

Analog VLSI design contributes the VLSI design and technology by the implementation of analog amplification circuits with respect to digital integration.

Analog Circuits such as Common Source amplifier, Common Gate amplifiers etc are implemented with respect to digital integration for amplification.

- On an observance analog contributes 20% and digital contributes 80%.

So, here we implement the analog circuits into CMOS transistor logic, for going digital integration, thus also studying there small signal model[1] becomes necessary.

**Small Signal Models of a MOSFET:** *(Considering nMOSFET)*

There are two types of small signal model of MOSFET i.e.:

1. Low Frequency Small Signal Model of MOSFET
2. High Frequency Small Signal Model of MOSFET

**Low Frequency Small Signal Model of MOSFET:**

In analog circuit, MOSFET devices are normally operated in saturation, here the drain current $I_{ds}$ of nMOSFET can be written as,

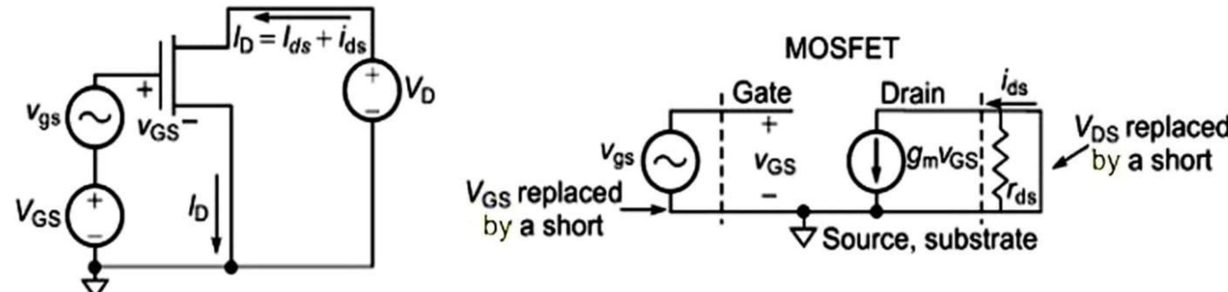
$$I_{ds} = \frac{\beta}{2}\left(v_{gs} - V_{THN}\right)^2\left[1 + \lambda_c\left(V_{ds} - V_{ds,sat}\right)\right]$$

Where:

$V_{ds,sat} \ll V_{ds}$

$\lambda_c$ is Channel length modulation parameter

$\beta = \frac{\mu c_{ox} W}{L}$ ($\mu$ is Electron Mobility, $c_{ox}$ is Gate oxide capacitance, $W$ is Width and $L$ is the length)




*CIRCUIT OF MOSFET AND ITS LOW FREQUENCY SMALL SIGNAL MODEL*

[1] ***Small signal models are used to analyze transistor circuits easily and rapidly. A small signal model is drawn by using simple approximations by retaining its essential features and discarding its less important features***

The circuit diagram of a nMOSFET and its small signal model is shown above, the dc sources are labelled say as $V_{GS}$ and ac sources with their subscripts i.e. $v_{gs}$ respectively.

Assuming $V_{GS} >> v_{gs}$. Since the MOSFET is in the saturation region, $V_{DS} > V_{GS} - V_{THN}$, the total (ac + dc) drain current is given as,

$$I_D = I_{ds} + i_{ds} = \frac{\beta}{2}\left(V_{GS} + v_{gs} - V_{THN}\right)^2 (1 + \lambda_c V_{ds})$$

The forward transconductance, $g_m$, of the MOSFET is given as,

$$g_m = \left.\frac{\partial I_D}{\partial v_{GS}}\right|_{V_{GS}=Constant} = \beta\left(V_{GS} + v_{gs} - V_{THN}\right)(1 + \lambda_c V_{ds})$$

For the small signal low frequency ac equivalent circuit, it is seen that $V_{ds}$ and $V_{GS}$ are replaced by a short and for small signal, $v_{gs} << V_{GS}$. Since $\lambda_c < 1$ and $\lambda_c V_{ds} << 1$, hence the transconductance can be written as,

$$\boldsymbol{g_m = \beta(V_{GS} - V_{THN}) = \sqrt{2\beta I_{ds}}}$$

The output resistance is given as,

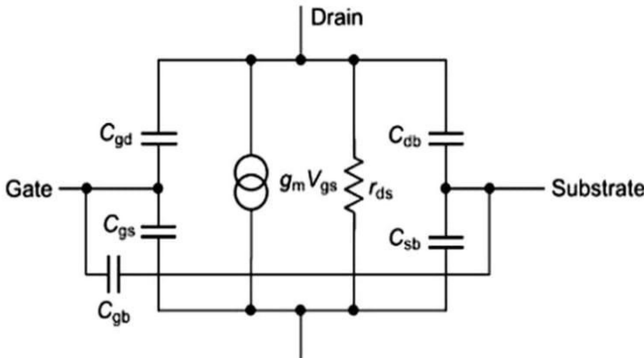
$$\boldsymbol{r_{out} = r_{ds} = \left.\frac{\partial V_{ds}}{\partial I_D}\right|_{V_{GS}=Constant} = \frac{1}{\beta(V_{GS} - V_{THN})^2 \lambda_c} = \frac{1}{\lambda_c I_{ds}}}$$

The maximum voltage gain can be given as,

$$\boldsymbol{A_V = \frac{\partial V_{ds}}{\partial V_{gs}} = -\frac{1 \Big/ \frac{\partial V_D}{\partial V_{ds}}}{1 \Big/ \frac{\partial I_D}{\partial V_{gs}}} = -\frac{\sqrt{2\beta I_{ds}}}{\lambda_c I_{ds}} = -\frac{\sqrt{2\beta}}{\lambda_c \sqrt{I_{ds}}}}$$

## High Frequency Small Signal Model of MOSFET:

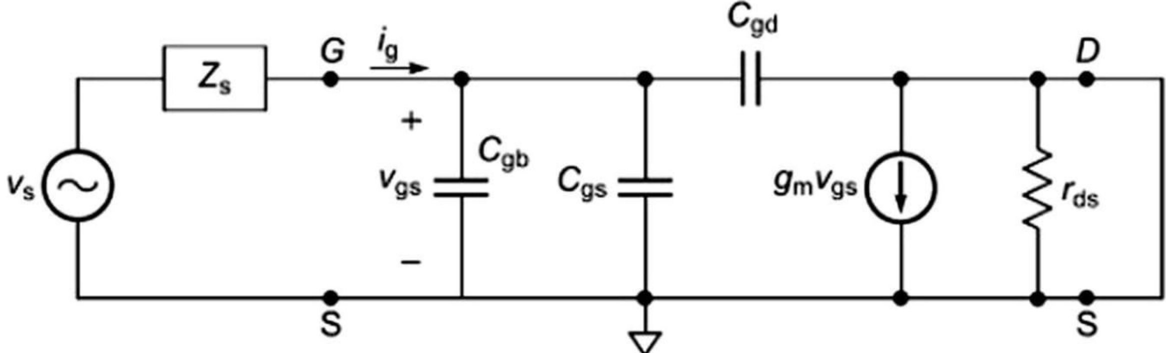
In the high frequency model of MOSFET, the MOSFET capacitances are considered as shown.




*HIGH FREQUENCY SMALL SIGNAL MODEL*

As seen the capacitances of drain diffusion region, source diffusion regions and gate over field region are denoted by $C_{db}$, $C_{sb}$ and $C_{gb}$. The capacitances of gate- drain and gate- source are denoted by $C_{gd}$ and $C_{gs}$.

The above high frequency small signal model can be further simplified as shown below,




*SIMPLIFIED HIGH FREQUENCY SMALL SIGNAL MODEL*

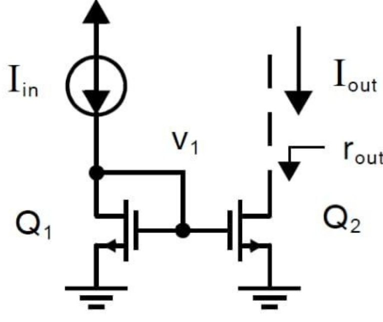
The current gain can be given as,

$$\frac{i_{ds}}{i_g} = \frac{\beta(V_{gs} - V_{THN})}{j2\pi f(C_{gb} + C_{gs} + C_{gd})}$$

**What is a Current Mirror/ Ideal Current Mirror?**

An ideal current mirror is a two-port circuit that accepts an input current $I_{in}$ and produces and output current $I_{out} = I_{in}$. Since current sensing is best done with a low resistance, for example an ammeter, the ideal current source will have zero input resistance. An ideal current source has a high output resistance and, hence, so will an ideal current mirror. In this way, the ideal current mirror faithfully reproduces the input current regardless of the source and load impedances to which it is connected.

**<u>Simple CMOS Current Mirror</u>:**




*A SIMPLE CMOS CURRENT MIRROR*

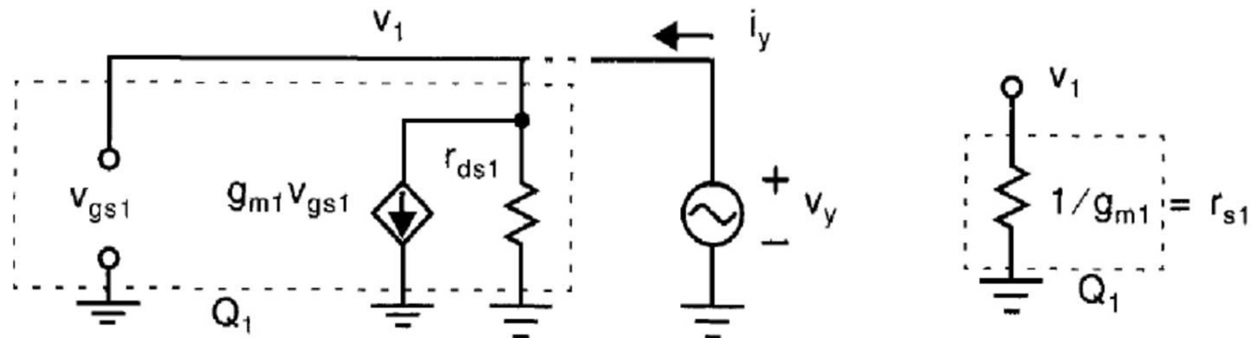
A simple CMOS current mirror is shown in figure, in which it is assumed that both the transistors are in the active region. If the finite output impedances of the transistors ($Q_1$ and $Q_2$) are ignored and are of same size, then $Q_1$ and $Q_2$ will have the same current since both have same gate- source voltage $V_{gs}$.

When considering finite output impedance, which ever transistor has a larger drain- source voltage $V_{ds}$ will also have a larger current and the finite output impedance of the transistors will cause the small signal output impedance of the current mirror i.e. the small signal impedance looking into the drain of $Q_2$, to be less than infinite.

To find the output impedance of the current mirror, $r_{out}$, the small signal circuit is analyzed after placing a signal source $_X$, at the output node. As per definition,

$$r_{out} = \frac{V_X}{i_X}$$

Where $i_X$ is the current flowing out of the source and into the drain of $Q_2$.




*SMALL SIGNAL MODEL FOR $Q_1$* *EQUIVALENT SMALL SIGNAL MODEL FOR $Q_1$*

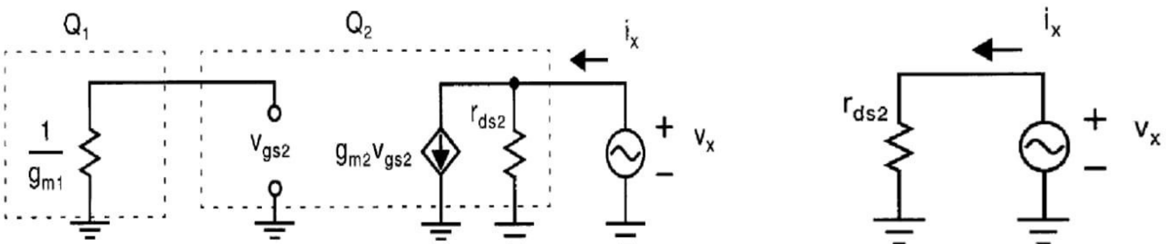
Consider the small signal model of $Q_1$ alone as shown which is a low frequency small signal model. It is to be noted that $Q_1$ is diode connected i.e. its drain and gate are connected and $I_{in}$ does not exist in the small signal model, $I_{in}$ is replaced with an open circuit because it is an independent current source. This model can further be reduced by Thevenins equivalent circuit. The Thevenins equivalent output voltage is 0 since the circuit is stable and contains no input signal. This circuit's Thevenins equivalent output impedance is found by applying a test signal voltage, $V_y$ at $V_1$ and measuring the signal current, $i_Y$, as shown. Here the current $i_y$ is given as,

$$\boldsymbol{i_y = \frac{V_y}{r_{ds1}} + g_{m1}V_{gs1} = \frac{V_y}{r_{ds1}} + g_{m1}V_y}$$

Recalling the output impedance is given by,

$$r_{out} = \frac{V_y}{i_y} = \frac{1}{g_{m1}} \,||\, r_{ds1}$$

This is because $r_{ds1} >> 1/ g_{m1}$. We approximate the output impedance equal to $1/ g_{m1}$, which results in the equivalent model as shown for $Q_1$.

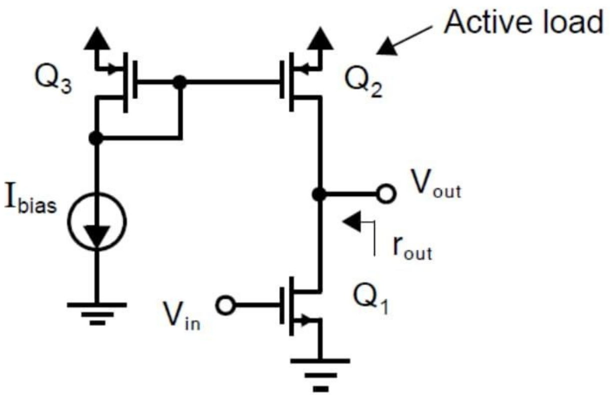



SMALL SIGNAL MODEL FOR CMOS CURRENT MIRROR EQUIVALENT SMALL SIGNAL MODEL FOR CMOS CURRENT MIRROR

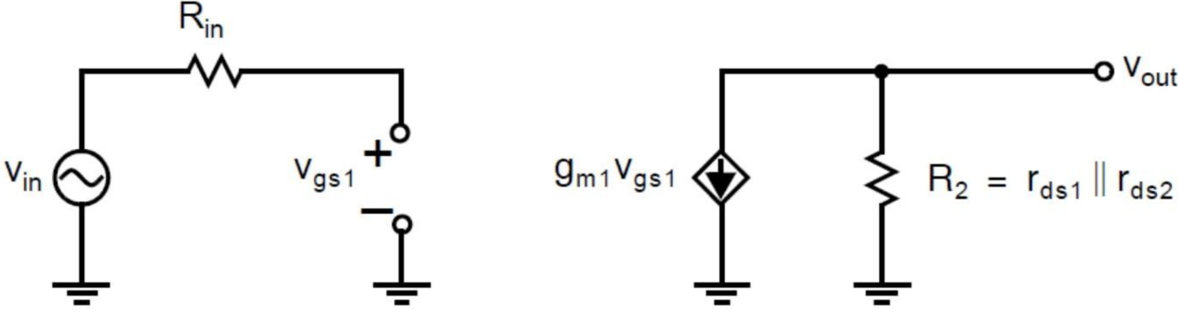
Further we lead to the small signal model of the CMOS current mirror as shown, where $V_{gs2}$ has been connected to ground via a resistance of 1/ $g_{m1}$. Since no current flows through 1/ $g_{m1}$ resistor, $V_{gs2} = 0$, no matter what voltage $V_x$ is applied to the current mirror output. This is no surprise since MOS transistors operate unilaterally at low frequencies. Thus, since $g_{m2}\ V_{gs2} = 0$, the circuit is simplified to the equivalent small signal model as shown. The small signal output impedance, $r_{out}$ is equal to $r_{ds2}$.

*Therefore, a simple CMOS current mirror has a small-signal input resistance of* 1/ $g_{m1}$ *and a small-signal output resistance* $r_{ds2}$.

**Common Source Amplifier:**




A COMMON SOURCE AMPLIFIER




SMALL SIGNAL EQUIVALENT MODEL FOR COMMON SOURCE AMPLIFIER

The common-source amplifier is a popular gain stage, especially when high input impedance is desired as shown. The use of an active load takes advantage of the nonlinear, large-signal transistor current–voltage relationship to provide large small-signal resistances without large dc voltage drops.

Here, an n- channel common-source amplifier has a p- channel current mirror used as an active load to supply the bias current for the drive transistor. By using an active load, a high- impedance output load can be realized without using excessively large resistors or a large power- supply voltage. As a result, for a given power supply voltage, a larger voltage gain can be achieved using an active load than would be possible if a resistor were used for the load.

For example, if a 100 kΩ load were required with a bias current, a resistive- load approach would require a power- supply voltage of 100 kΩ × 100 μA. An active load takes advantage of the nonlinear, large-signal transistor current voltage relationship to provide large small-signal resistances without large dc voltage drops.

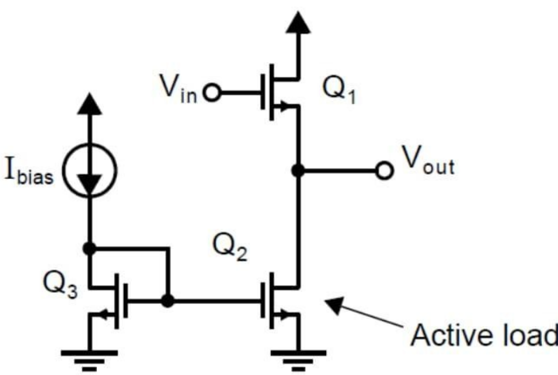
The small- signal equivalent circuit for the low-frequency analysis of the common-source amplifier as shown, where $V_{in}$ and $R_{in}$ are the Thévenin equivalent of the input source. It is assumed that the bias voltages are such that both transistors are in the active region. The output resistance, $R_2$- is made up of the parallel combination of the drain- to- source resistance of $Q_1$, i.e. $r_{ds1}$, and the drain- to- source resistance of $Q_2$, i.e. $r_{ds2}$. Notice that the voltage- controlled current source modeling the body effect has not been included since the source is at a small signal ground, and, therefore, this source always has 0 current.

Using the small signal analysis, we have $v_{gs1} = v_{in}$ and, therefore,

$$A_V = \frac{V_{out}}{V_{in}} = -g_{m1}R_2 = -g_{m1}(r_{ds1} \| r_{ds2})$$

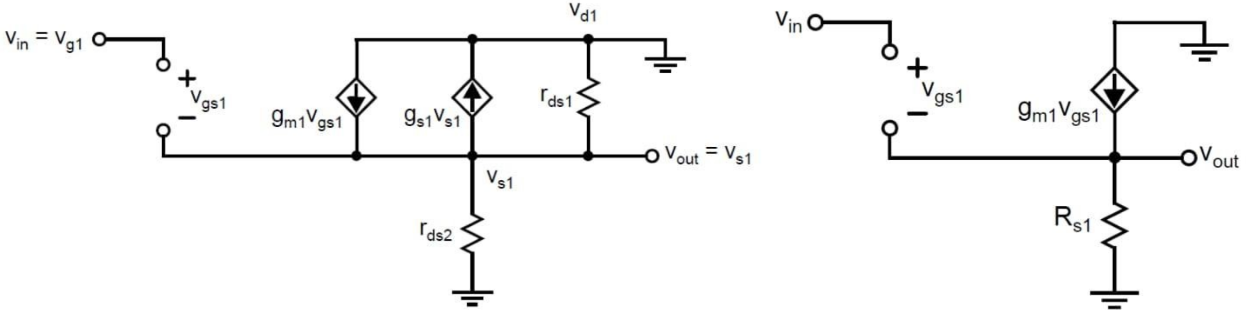
Depending on the device sizes, currents, and the technology used, a typical gain for this circuit is in the range of -5 to -100. To achieve similar gains with resistive loads, much larger power-supply voltages must be used which also greatly increases the power dissipation.

## <u>Common Drain Amplifier or Source Follower</u>:




*A COMMON DRAIN AMPLIFIER OR SOURCE FOLLOWER*

A general use of current mirrors is to supply the bias current of source- follower amplifiers, as shown. Here $Q_1$ is the source follower and $Q_2$ is an active load that supplies the bias current of $Q_1$. These amplifiers are commonly used as *voltage buffers* and are therefore commonly called source followers.
They are also referred to as common-drain amplifiers, since the input and output nodes are at the gate and source nodes respectively, with the drain node being at small-signal ground. Although the dc level of the output voltage is not the same as the dc level of the input voltage, ideally the small-signal voltage gain is close to unity. In reality, it is somewhat less than unity. However, although this circuit does not generate voltage gain, it does have the ability to generate current gain.




*LOW FREQUENCY SMALL SIGNAL MODEL* *EQUIVALENT SMALL SIGNAL MODEL*

It is to be noted that the voltage-controlled current source that models the body effect of MOS transistors has been included because the source is not at small- signal ground. The body effect is a major limitation on the small signal gain. From the small signal it is seen that $r_{ds1}$ is parallel to $r_{ds2}$, also that the voltage-controlled current source modelling the body effect produces a current that is proportional to the voltage across it. This relationship makes the body effect equivalent to a resistor of size $1/\ g_{s1}$, which is also in parallel with $r_{ds1}$ and $r_{ds2}$. Thus the small signal model is reduced to equivalent small signal model as shown in which $R_{s1} = r_{ds1} \parallel r_{ds2} \parallel 1/\ g_{s1}$.

Now writing the nodal equation at $V_{out}$, and noting that $V_{gs1} = V_{in} - V_{out}$, we have,

$$V_{out}/R_{s1} - g_{m1}(V_{in} - V_{out}) = 0$$

To minimize circuit equation errors, a consistent methodology should be maintained when writing nodal equations. The methodology employed here is as follows: The first term is always the node at which the currents are being summed. This node voltage is multiplied by the sum of all admittances connected to the node. The next negative terms are the adjacent node voltages, and each is multiplied by the connecting admittance. The last terms are any current sources with a multiplying negative sign used if the current is shown to flow into the node.

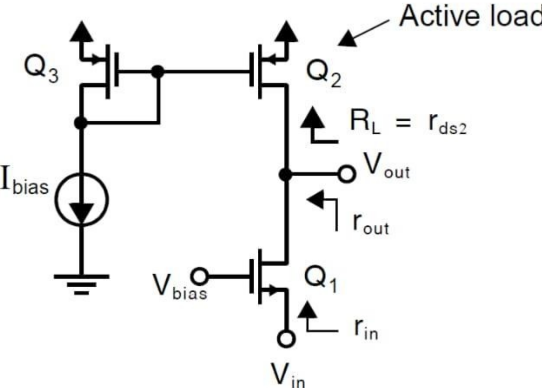
Solving for $V_{out}/\ V_{in}$, we have,

$$A_v = \frac{V_{out}}{V_{in}} = \frac{g_{m1}}{g_{m1} + G_{s1}} = \frac{g_{m1}}{g_{m1} + g_{s1} + g_{ds1} + g_{ds2}} = g_{m1}\left(\frac{1}{g_{m1}} \parallel \frac{1}{g_{s1}} \parallel r_{ds1} \parallel r_{ds2}\right)$$

Normally, $g_{s1}$ is on the order of one-tenth to one-fifth that of $g_{m1}$. Also, the transistor output admittances, $g_{ds1}$ and $g_{ds2}$, might be one-tenth that of the body-effect parameter, $g_{s1}$. Therefore, it is seen that the body-effect parameter is the major source of error causing the gain to be less than unity. Notice also that at low frequencies the stage is completely unilateral. In other words, there is no signal flow from the output to the input.

**Common Gate Amplifier:**

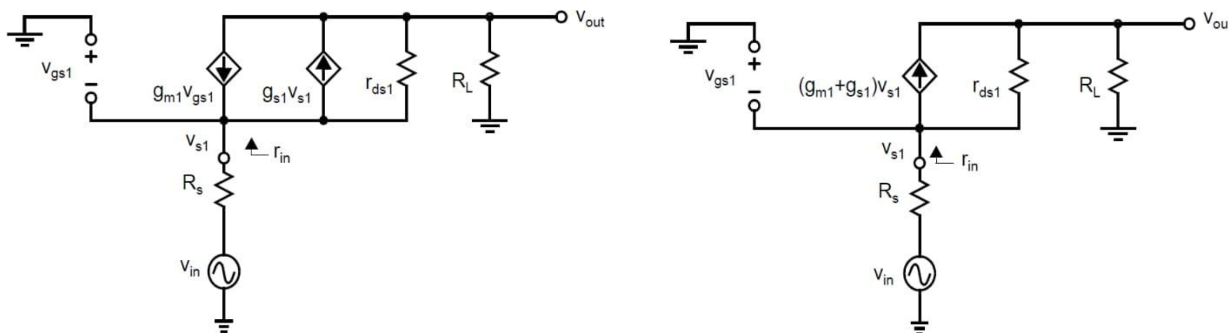
The common-gate amplifier as shown provides a voltage gain comparable to that of the common-source amplifier, but with a relatively low input resistance on the order of $1/g_m$.




*A COMMON GATE AMPLIFIER- WITH CURRENT MIRROR ACTIVE LOAD*

A common application for a common- gate amplifier is as the first stage of an amplifier where the input signal is a current, in such cases a small input impedance is desired in order to ensure all of the current signal is drawn into the amplifier, and none is "lost" in the signal source impedance. Aside from its low input impedance, the common gate amplifier is similar to a common-source amplifier, in both cases the input is applied across $V_{gs}$, except with opposite polarities, and the output is taken at the drain. Hence, in both cases the small signal gain magnitude approximately equals the product of $g_m$ and the total impedance at the drain.

If we use straightforward small-signal analysis, when the impedance seen at (in this case, the output impedance of the current mirror formed by $Q_2$) is much less than $r_{ds1}$, the input impedance, $r_{out}$, is found to be $1/\ g_{m1}$ at low frequencies. However, in integrated applications, the impedance seen at $V_{out}$ is often on the same order of magnitude or even much greater than $r_{ds1}$. In this case, the input impedance at low frequencies can be considerably larger than $1/\ g_{m1}$. To see this result, consider the small-signal model as shown. In this model, the voltage-dependent current source that models the body effect has been included. It is noticed that $V_{gs1} = -\ V_{s1}$ and therefore the two current sources can be combined into a single current source, as shown in the simplified small signal model. This simplification is always possible for a transistor that has a grounded gate in a small-signal model, and considerably simplifies taking the body effect into account.




*LOW FREQUENCY SMALL SIGNAL MODEL*

*EQUIVALENT OR SIMPLIFIED SMALL SIGNAL MODEL*

The body effect for transistors with grounded gates can be ignored, and then, after the analysis is complete, simply replace the constants $g_{mi}$ with $g_{mi} + g_{si}$. However, now we include the body-effect parameter throughout the analysis.

At node $V_{out}$, we have

$$v_{out}(G_L + g_{ds1}) - v_{s1}g_{ds1} - (g_{m1} + g_{s1})v_{s1} = 0 \quad \text{(Equation 1)}$$

Rearranging slightly, we have,

$$\frac{v_{out}}{v_{s1}} = \frac{g_{m1} + g_{s1} + g_{ds1}}{G_L + g_{ds1}} = (g_{m1} + g_{s1} + g_{ds1})(R_L \| r_{ds1}) \cong g_{m1}(R_L \| r_{ds1}) \quad \text{(Equation 2)}$$

Here the gain is approximately equal to ${g_{m1}}/{(G_L + g_{ds1})}$.

The current going into the source of $Q_1$ is given by,

$$i_s = v_{s1}(g_{m1} + g_{s1} + g_{ds1}) - v_{out}g_{ds1} \quad \text{(Equation 3)}$$

Combining equations 2 and 3 to find the input admittance $y_{in} = {1}/{r_{in}}$ we have,

$$y_{in} = \frac{i_s}{v_{s1}} = \frac{g_{m1}+g_{s1}+g_{ds1}}{1+\frac{g_{ds1}}{G_L}} = \frac{g_{m1}}{1+\frac{g_{ds1}}{G_L}} \quad \text{(Equation 4)}$$

Alternatively, we have,

$$r_{in} = \frac{1}{y_{in}} = \frac{1}{g_{m1}}\left(1 + \frac{R_L}{r_{ds1}}\right)$$

With the p- channel active load, $R_L = r_{ds2}$. Since, in this case, is approximately the same magnitude as $r_{ds1}$, the input impedance, $r_{in}$, is about $2/\ g_{m1}$for low frequencies- twice as large as the expected value of $1/\ g_{m1}$. This increased input impedance must be taken into account in applications such as transmission- line terminations. In some examples, the current-mirror output impedance realized by is much larger than (i.e. $R_L >> r_{ds1}$, ), and so the input impedance for this common-gate amplifier is much larger than $1/\ g_{m1}$. This increased input impedance often occurs in integrated circuits and is not commonly known.

The attenuation from the input source to the transistor source can be considerable for a common-gate amplifier when $R_s$ is large. This attenuation is given by,

$$\frac{V_{s1}}{V_{in}} = \frac{G_s}{G_s + y_{in}}$$

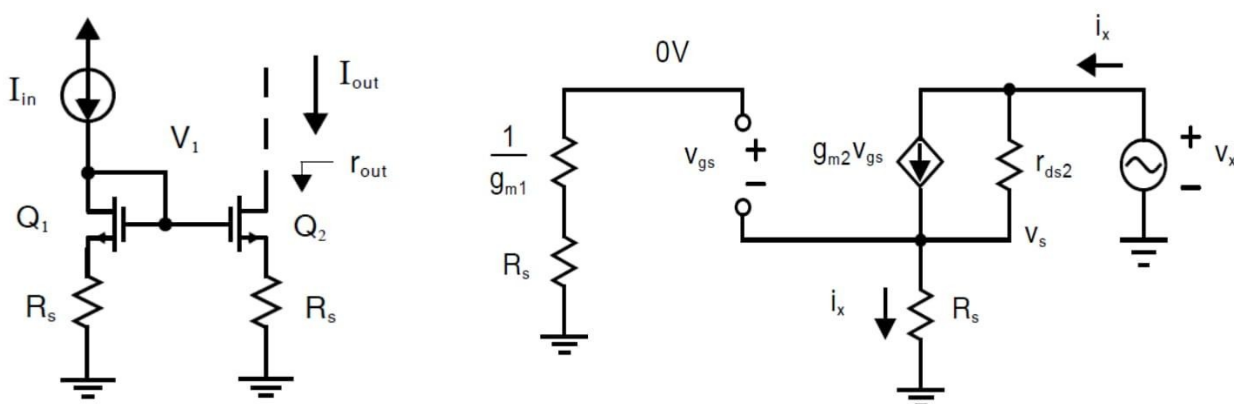
Using equation 4 to replace $y_{in}$, we have- using admittance- divider rule,

$$\frac{v_{s1}}{V_{in}} = \frac{G_s}{G_s+\frac{g_{m1}}{1+\frac{g_{ds1}}{G_L}}} \quad \text{(Equation 5)}$$

Using equation 2 and 5, to find the overall dc gain- which is given as,

$$A_V = \frac{V_{out}}{V_{in}} = \left[\frac{G_s}{\left(G_s + \frac{g_{m1}}{1+\frac{g_{ds1}}{G_L}}\right)}\right]\frac{g_{m1}}{G_L + g_{ds1}}$$

## Source Degenerated Current Mirror:




*SOURCE DEGENERATED CURRENT MIRROR AND ITS SMALL SIGNAL MODEL*

A source degenerated current mirror is used to increase the output impedance. A source degenerated current mirror is shown along with its small signal model.

Here since no current flow through the gate, the gate voltage is 0 V. The current $i_x$ sourced by the applied voltage source is equal to the current through the degeneration resistor $R_s$. Therefore, we have,

$$V_s = i_x R_s \quad \text{(Equation 1)}$$

And

$$V_{gs} = -V_s \quad \text{(Equation 2)}$$

Setting $i_x$ equal to the total current through $g_{m2}$ $V_{gs}$ and $r_{ds2}$ gives,

$$i_x = g_{m2}V_{gs} + \frac{V_x - V_s}{r_{ds2}} \quad \text{(Equation 3)}$$

Substituting equation 1 and 2 in 3 gives,

$$i_x = -i_x g_{m2} R_s + \frac{v_x - i_x R_s}{r_{ds2}}$$ (Equation 4)

Rearranging, we find the output impedance to be given as,

$$r_{out} = \frac{v_x}{i_x} = r_{ds2}[1 + R_s(g_{m2} + g_{ds2})] \cong r_{ds2}(1 + R_s g_{m2})$$ (Equation 5)

Where $g_{ds2}$ is equal to $1/r_{ds2}$, which is much less than $g_{m2}$.

Since the gate is at a small signal ground, the body effect can be considered by replacing $g_{m2}$ in the above equation with $g_{m2} + g_{s2}$.

Therefore equation 5 becomes,

$$r_{out} = \frac{v_x}{i_x} = r_{ds2}[1 + R_s(g_{m2} + g_{s2} + g_{ds2})] \cong r_{ds2}[1 + R_s(g_{m2} + g_{s2})]$$

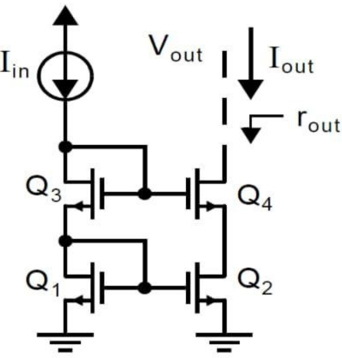
***HIGH OUTPUT IMPEDANCE CURRENT MIRRORS:***

The current mirrors which have output impedances that are larger than that of a simple current mirror by a factor of $g_m r_{ds}$- the maximum gain of a single transistor are **high output impedance current mirrors**.

The two types of high output impedance current mirrors are:

1. Cascode Current Mirror
2. Wilson Current Mirror

**Cascode Current Mirror:**




*CASCODE CURRENT MIRROR*

A cascode current mirror is shown in figure, where the output impedance looking into the drain of $Q_2$ is $r_{ds2}$, which is analysed very similar to the analysis of a simple current mirror. Therefore the output impedance can be immediately derived by considering $Q_4$ as a current source with a source degeneration resistor of value equal to $r_{ds2}$.

It is to be noted that the addition of a cascode device to a CMOS current mirror increases its output resistance by approximately the gain of the cascode device, $g_m\, r_{ds}$.
Using,

$$r_{out} = \frac{v_x}{i_x} = r_{ds2}[1 + R_s(g_{m2} + g_{ds2})] \cong r_{ds2}(1 + R_s g_{m2})$$

And noting that $Q_4$ is now the cascode transistor rather than $Q_2$, we have,

$$r_{out} = r_{ds4}[1 + R_s(g_{m4} + g_{s4} + g_{ds4})]$$

Where now $Rs = r_{ds2}$. Therefore the output impedance is given as,

$$\begin{aligned} r_{out} &= r_{ds4}[1 + r_{ds2}(g_{m4} + g_{s4} + g_{ds4})] \\ &\cong r_{ds4}[1 + r_{ds2}(g_{m4} + g_{s4})] \\ &\cong r_{ds4}(r_{ds2} g_{m4}) \end{aligned}$$

Thus, the output impedance has been increased by a factor of $g_m\, r_{ds2}$, which is an upper limit on the gain of a single transistor MOS gain- stage, and might be a value between 10 and 100, depending on the transistor sizes and currents and the technology being used. This significant increase in output impedance can be instrumental in realizing single stage amplifiers with large low frequency gains.

The disadvantage in a cascode current mirror is that it reduces the maximum output signal swings possible before transistors enter the triode region. This can be illustrated by having a n- channel transistor to be in the active region- also called the saturation or pinch- off region, its drain- source voltage must be greater than $V_{eff}$, where $V_{eff}$ is,

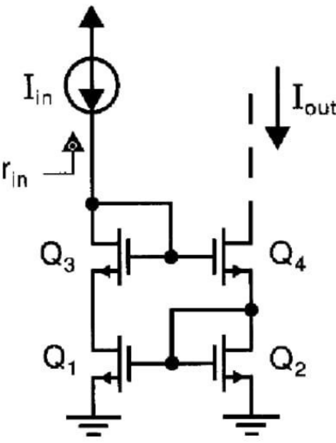
$$V_{eff} \equiv V_{GS} - V_{tn}$$

Which is further given by,

$$V_{eff} = \sqrt{\frac{2I_D}{\mu_n C_{ox}(W/L)}}$$

## Wilson Current Mirror:

The Wilson current mirror is an example of using shunt- series feedback to increase the output impedance. The Wilson current mirror is shown in figure.




*WILSON CURRENT MIRROR*

Here $Q_2$ senses the output current and then the mirrors it to $I_{D1}$, which in turn is subtracted from the input current $I_{in}$, it is to be noted that $I_{D1}$ must be precisely equal $I_{in}$- otherwise the voltage at the gate of $Q_3$, $Q_4$ would either increase or decrease, and the negative feedback loop forces this equality. This feedback arrangement increases the output impedance by an amount equal to 1 plus the loop gain.

Assuming all devices are matched, the output impedance without the feedback due to $Q_1$, $Q_3$ would be 2 $r_{ds4}$, taking into account that $Q_4$ has source degeneration equal to 1/ $g_{m2}$ (i.e. the small signal impedance of diode connected $Q_2$), which is responsible for the 2 factor.

The loop gain is approximately given as,

$$A_L = \frac{g_{m1}(r_{ds1} || r_{in})}{2}$$

Where $r_{in}$ is the input impedance of the biasing current source $I_{in}$.

The factor of 1/ 2 is due to the voltage attenuation from the gate of $Q_4$ to its source, caused by the source degeneration of the diode connected $Q_2$.

Assuming $r_{in}$ is approximately equal to $r_{ds1}$, then the loop gain is given by,

$$A_L = \frac{g_{m1} r_{ds1}}{4}$$

And the output impedance is therefore given as,

$$r_{out} = r_{ds4} \frac{g_{m1}(r_{ds1} || r_{in})}{2} = r_{ds4} \frac{g_{m1} r_{ds1}}{2}$$

It is seen from the equation of output impedance that the output impedance of the Wilson current mirror is roughly one- half the output impedance for that of a cascode current mirror. For this reason the cascode current mirror is often preferred over the Wilson current mirror.

The output voltage swing, the minimum allowed voltage across the current mirror before $Q_4$ enters the triode region is 2 $V_{eff1}$ + $V_{in}$, which is similar to that of the cascode current mirror.

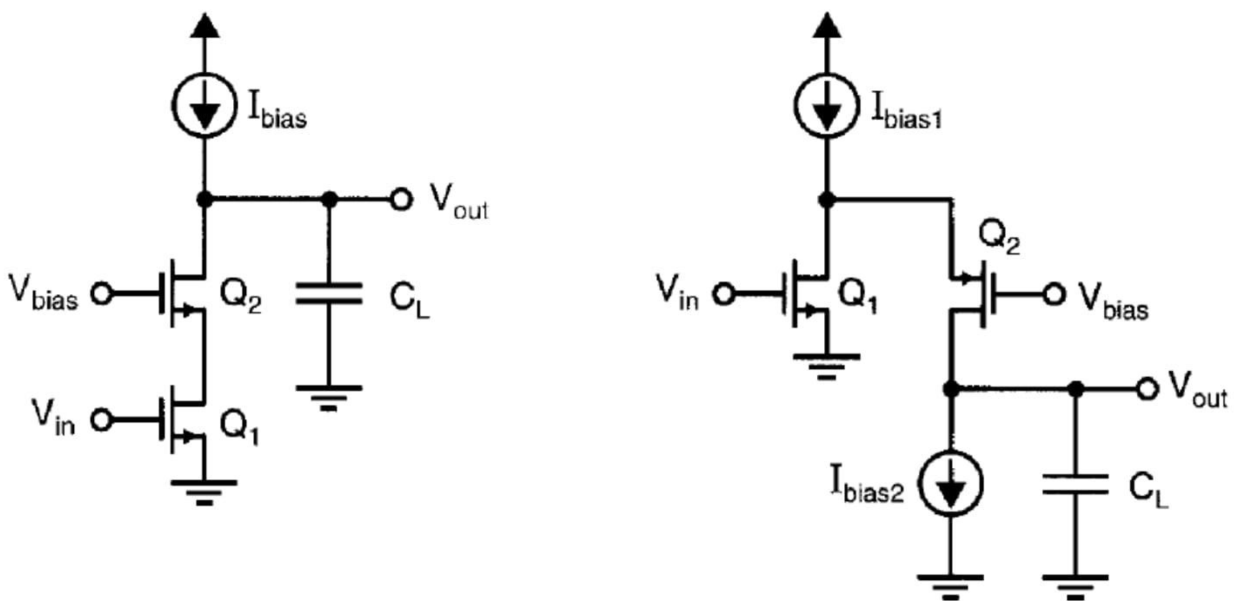
Therefore, it is noted that in the $Q_3$ is not required in the Wilson current mirror, it has been included to give $Q_1$ and $Q_2$ the same drain- source bias voltages and thus minimizes the inaccuracies caused by the large signal output impedances of the transistors. Without this transistor the output current would be slightly smaller than the input current because $V_{DS1}$ would be larger than $V_{DS2}$, keeping the small signal output impedance remaining the same.

## Cascode Amplifier or Cascode Gain Stage:

The cascode amplifier is commonly used configuration in digital IC design which consists of a common source connected transistor feeding into a common gate connected transistor.

Two illustrations/ types of cascode amplifiers are:

1. Telescopic cascode amplifier
2. Folded- cascode amplifier




TELESCOPIC CASCODE AMPLIFIER FOLDED CASCODE AMPLIFIER

As seen telescopic cascode amplifier has both common n- channel transistors $Q_1$- common source and $Q_2$- common gate, this is referred as a *telescopic cascode amplifier* and as seen a folded cascode amplifier has a n- channel common source transistor $Q_1$ and a p- channel common gate , this is referred as a *folded cascode amplifier*.

The folded cascode amplifier allows the dc level of the output signal to be the same as the dc level of the input signal, but this folded cascode amplifier is slower than the telescopic cascode amplifier because of the impedance levels of the folded cascode is roughly three times larger due to the smaller transconductance of the p- channel transistors as compared to n- channel transistors, although parasitic capacitances at the source of the cascode transistor are similar in both the amplifier cases.

There are two major reasons for the demand of these amplifier stages i.e.:

1. They can have larger gain for a single stage due to large impedances at the output- to enable this high gain the current sources connected to the output nodes are realized using high quality cascode current mirrors. This gain is obtained without degradation in speed and sometimes also with an increase in speed.
2. They limit the voltage across the input drive transistor- this minimizes any shorted short channel effects which becomes more important with modern technologies having very low short channel length transistors.

The analysis of the cascode gain stage/ amplifier is based on the telescopic stage,

WKT, using current mirrors, the impedance looking into the drain of cascode transistor $Q_2$ is approximately given by,

$$r_{d2} = g_{m2}\, r_{ds1}\, r_{ds2}$$

The total impedance at the output node is $r_{ds2}$ in parallel with $R_L$, where $R_L$ is the output impedance of the bias current source, $I_{bias}$. Assuming $I_{bias}$ is a high quality source with output impedance of the order of,

$$R_L = g_{m-p}\, r^2{}_{ds-p} \qquad \text{(Equation 1)}$$

Then the total impedance at the output node is,

$$R_{out} = \frac{g_m\, r^2{}_{ds}}{2} \qquad \text{(Equation 2)}$$

Now, to find the approximate low- frequency gain, we use a part of the analysis of the common gate amplifier. i.e.

$$y_{in} = \frac{i_s}{v_{s1}} = \frac{g_{m1} + g_{s1} + g_{ds1}}{1 + \frac{g_{ds1}}{G_L}} = \frac{g_{m1}}{1 + \frac{g_{ds1}}{G_L}}$$

Therefore, the low frequency impedance looking into the source of the common gate or cascode, transistor $Q_2$ is given as,

$$y_{in2} = \frac{g_{m2} + g_{s2} + g_{ds2}}{1 + \frac{g_{ds2}}{G_L}} = \frac{g_{m2}}{1 + \frac{g_{ds2}}{G_L}} \qquad \text{(Equation 3)}$$

Substituting equation 1 in equation 3, we get,

$$y_{in\,2} = \frac{g_m}{1 + \frac{g_{ds}}{g^2{}_{ds}/g_m}} = g_{ds} \quad \text{(Equation 3)}$$

Therefore, the gain from the input to the source of $Q_2$ is given by,

$$\frac{v_{s2}}{v_{in}} = \frac{g_{m1}}{g_{ds1} + y_{in2}} = -\frac{g_m}{2g_{ds}}$$

Therefore, the overall gain is given by,

$$A_V = \frac{v_{s2}}{v_{in}}\frac{v_{out}}{v_{s2}} = -\frac{g_m}{2g_{ds}}\frac{g_{m2}}{G_L + g_{ds2}} = -\frac{g_m}{2g_{ds}}\frac{g_{m2}}{g_{ds2}} = -\frac{1}{2}\left(\frac{g_m}{g_{ds}}\right)^2$$

**Problems:**

1. Considering the current mirror shown, where $I_{in}$ = 100 μA and each transistor has W/L = 100 μm/ 1.6 μm. Given that $\mu_n C_{ox}$ = 92 μA/ $V^2$, $V_{tn}$ = 0.8V and $r_{ds}$ = [8000L (μm)] / [$I_D$ (mA)], find $r_{out}$ for the current mirror and the value of $g_{m1}$. Also, estimate the change in $I_{out}$ for 0.5V change in the output voltage.

Sol.

Given:

$I_{in}$ = 100 μA

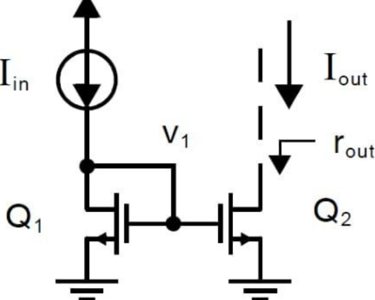
W/L = 100 μm/ 1.6 μm

$\mu_n C_{ox}$ = 92 μA/ $V^2$

$V_{tn}$ = 0.8V

$r_{ds}$ = [8000L (μm)] / [$I_D$ (mA)]

ΔV = 0.5V




To find:

$r_{out}$ = ?

$g_{m1}$ = ?

$\Delta I_{out}$ = ?

Here, W/L ratios of $Q_1$ and $Q_2$ are same, the nominal value of $I_{out}$ equals that of $I_{in}$ = 100 µA.

Therefore, $r_{out}$ is given as,

$$r_{out} = r_{ds2} = \frac{8000\ L\ (\mu m)}{I_D\ (mA)} = \frac{8000 \times 1.6}{0.1} = 128\ k\Omega$$

The value of $g_{m1}$ is given by,

$$g_{m1} = \sqrt{2\mu_n C_{ox}\ (W/L)\ I_{D1}} = \sqrt{2 \times 92 \left(\frac{100}{1.6}\right) \times 100} = 1.07\ mA/\ V$$

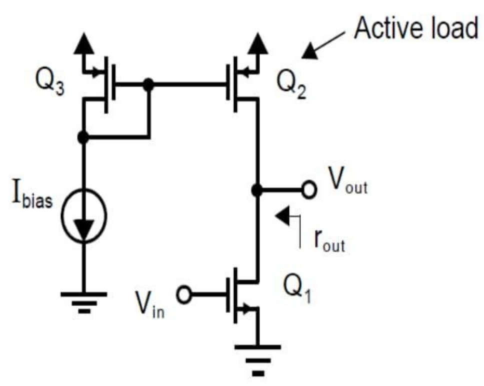
The change in the output current can be estimated as,

$$\Delta I_{out} = \frac{\Delta V}{r_{out}} = \frac{0.5}{128} = 3.9\ \mu A$$

2. Assuming all transistors have W/L = 100 µm/ 1.6 µm as shown with $\mu_n C_{ox}$ = 90 µA/ $V^2$, $\mu_p C_{ox}$ = 30 µA/ $V^2$, $I_{bias}$ =100 µA, $r_{ds-n}$ = [8000L (µm)] / [$I_D$ (mA)] and $r_{ds-p}$ = [12000L (µm)] / [$I_D$ (mA)]. Find the gain of the stage?

Sol.

Given:




W/L = 100 µm/ 1.6 µm

$\mu_n C_{ox}$ = 90 µA/ $V^2$

$\mu_p C_{ox}$ = 30 µA/ $V^2$

$I_{bias}$ =100 µA

$r_{ds-n}$ = [8000L (µm)] / [$I_D$ (mA)]

$r_{ds-p}$ = [12000L (µm)] / [$I_D$ (mA)]

To find:

$A_V$ = ?

WKT, the gain $A_V$ is given as,

$$A_V = -g_m (r_{ds1} || r_{ds2})$$

Now finding,

$$g_{m1} = \sqrt{2\mu_n C_{ox}\,(W/L)\,I_{bias}} = \sqrt{2 \times 90\left(\frac{100}{1.6}\right) \times 100} = 1.06\ \text{mA/ V}$$

$$r_{ds1} = \frac{8000\ L\ (\mu m)}{I_D\ (mA)} = \frac{8000 \times 1.6}{0.1} = 128\ k\Omega\ \text{ and } r_{ds2} = \frac{12000\ L\ (\mu m)}{I_D\ (mA)} = \frac{12000 \times 1.6}{0.1} = 192\ k\Omega$$

Now substituting the above found values in,

$$A_V = -g_m (r_{ds1} || r_{ds2})$$

We get,

$$A_V = -1.06(128\ ||\ 192) = -81.4$$

Therefore, the gain of the stage/ amplifier is – 81.4 respectively.

3. Consider a source follower as shown, where all transistors have W/L = 100 μm/ 1.6 μm with $\mu_n C_{ox}$ = 90 μA/ $V^2$, $\mu_p C_{ox}$ = 30 μA/ $V^2$, $I_{bias}$ =100 μA, $r_{ds-n}$ = [8000L (μm)] / [$I_D$ (mA)] and $\gamma_n = 0.5\ V^{1/2}$. Find the gain of the stage?

Sol.

Given:

W/L = 100 μm/ 1.6 μm

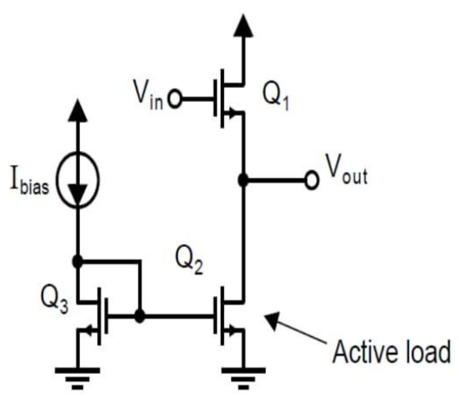
$\mu_n C_{ox}$ = 90 μA/ $V^2$

$\mu_p C_{ox}$ = 30 μA/ $V^2$

$I_{bias}$ =100 μA

$r_{ds-n}$ = [8000L (μm)] / [$I_D$ (mA)]

$\gamma_n = 0.5\ V^{1/2}$




To find:

$A_V$ = ?

WKT, the gain $A_V$ of a source follower is given as,

$$A_V = \frac{g_{m1}}{g_{m1} + g_{s1} + g_{ds1} + g_{ds2}} = \frac{g_{m1}}{g_{m1} + g_{s1} + {}^{1}/_{r_{ds1}} + {}^{1}/_{r_{ds2}}}$$

Now finding,

$$g_{m1} = \sqrt{2\mu_n C_{ox}\,(W/L)\,I_{bias}} = \sqrt{2 \times 90\left(\frac{100}{1.6}\right) \times 100} = 1.06\ mA/V$$

$$r_{ds1} = r_{ds2} = \frac{8000\ L\ (\mu m)}{I_D\ (mA)} = \frac{8000 \times 1.6}{0.1} = 128\ k\Omega$$

$$g_{s1} = \frac{\gamma\, g_m}{2\sqrt{V_{SB} + |2\phi_F|}} = \frac{0.5\ g_m}{2\sqrt{2 + 0.7}} = 0.15\, g_m = 0.15 \times 1.06 = 0.16\ mA/V$$

$$(\text{Here } V_{SB} = 2 \text{ and } \phi_F = 0.35 \rightarrow \text{Standard Values})$$

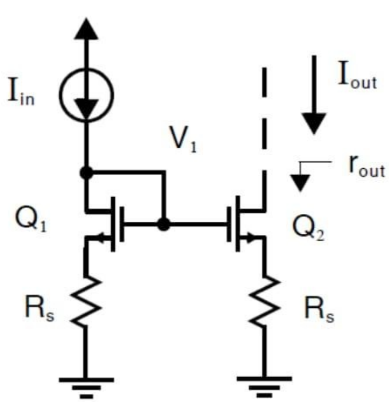
Therefore, the gain is given by,

$$A_V = \frac{1.06}{1.06 + 0.16 + {}^{1}/_{128} + {}^{1}/_{128}} = 0.86$$

Therefore, the gain of the source follower is 0.86 respectively.

**4.** Consider a source degenerated current mirror as shown, where $I_{in}$ = 100 µA, each transistor W/L = 100 µm/ 1.6 µm, $R_s$ = 5 kΩ, $\mu_n C_{ox}$ = 92 µA/ $V^2$, $V_{tn}$ = 0.8V and $r_{ds}$ = [8000L (µm)] / [$I_D$ (mA)]. Find $r_{out}$ for the current mirror. Assume the body effect can be approximated by $g_s = 0.2\ g_m$.

Sol.




Given:

W/L = 100 µm/ 1.6 µm

$R_s$ = 5 kΩ

$\mu_n C_{ox}$ = 92 µA/ $V^2$

$V_{tn}$ = 0.8V

$r_{ds}$ = [8000L (µm)] / [$I_D$ (mA)]

$g_s = 0.2\ g_m$

$I_{in}$ = 100 µA

To find:

$r_{out}$ = ?

WKT,

$$r_{out} = r_{ds2}[1 + R_s(g_{m2} + g_{s2})]$$

$I_{out} = I_{in}$

Now finding,

$$g_{m2} = \sqrt{2\mu_n C_{ox}\,(W/L)\,I_{out}} = \sqrt{2 \times 92\left(\frac{100}{1.6}\right) \times 100} = 1.07\ mA/V$$

$$r_{ds2} = \frac{8000\ L\ (\mu m)}{I_D\ (mA)} = \frac{8000 \times 1.6}{0.1} = 128\ k\Omega$$

Therefore, the $r_{out}$ is given by,

$$r_{out} = 128[1 + 5(1.07 + 0.2 \times g_m)]$$

$$r_{out} = 128[1 + 5(1.07 + 0.2 \times 1.07)] = 950\ k\Omega$$

Therefore, the output resistance $r_{out}$ of the source denegation current mirror is 950 kΩ respectively.

**5.** Consider a cascode current mirror as shown, where $I_{in}$ = 100 μA, each transistor W/L = 100 μm/ 1.6 μm, $\mu_n C_{ox}$ = 92 μA/ $V^2$, $V_{tn}$ = 0.8V and $r_{ds}$= [8000L (μm)] / [$I_D$ (mA)]. Find $r_{out}$ for the current mirror. Assume the body effect can be approximated by $g_s$ = 0.2 $g_m$. Also find the minimum output voltage at $V_{out}$ such that the output transistors remain in the active region.

Sol.

Given:

W/L = 100 μm/ 1.6 μm

$\mu_n C_{ox}$ = 92 μA/ $V^2$

$V_{tn}$ = 0.8V

$r_{ds}$ = [8000L (μm)] / [$I_D$ (mA)]

$g_s$ = 0.2 $g_m$

$I_{in}$ = 100 μA