CMR COLLEGE OF ENGINEERING & TECHNOLOGY
(AUTONOMOUS)
(A1425) VLSI DESIGN

VII - SEMESTER

L	T	P	C
4	1	0	4

Course coordinator Module coordinator

Course objective

The objective of the course is to

- Give exposure to different steps involved in the fabrication of ICs using MOS transistors, CMOS/BICMOS transistors and passive components
- Explain electrical properties of MOS and BICMOS devices to analyze the behaviour of inverters designed with various loads.
- Provide concept to design different types of logic gates using CMOS inverter and analyze their transfer characteristics.
- Provide design concepts to design building blocks of data path of any system using gates.
- Understand basic programmable logic devices and testing of CMOS circuits.

UNIT- I:

Introduction: Review of Semiconductors, Introduction to IC Technology, PMOS, NMOS, CMOS & BiCMOS technologies, Steps involved in Fabrication Process.

Basic Electrical Properties: Basic Electrical Properties of MOS and Bi CMOS Circuits: I_{ds}-V_{ds} relationships, MOS transistor threshold Voltage, Transconductance, figure of merit, NMOS Inverter, Various pull ups, CMOS Inverter analysis and design, Bi-CMOS Inverters.

UNIT- II:

VLSI Circuit Design Processes :VLSI Design Flow, Design Constraints, MOS Layers, Stick Diagrams, Design Rules and Layout, 2 um CMOS Design rules for wires, Contacts and Transistors Layout Diagrams for NMOS and CMOS Inverters and Gates, Scaling of MOS circuits, Limitations of Scaling.CAD tools

UNIT- III:

Gate Level Design: Logic Gates and Other complex gates, Switch logic, Alternate gate circuits, Basic circuit concepts, Concept of Sheet Resistance and Area Capacitance, Delays, Driving large Capacitive Loads, Wiring Capacitances, Fan-in and fan-out, Choice of layers.

UNIT- IV:

System Level Design Considerations: ALU unit, Multipliers, Parity generators, Comparators, Zero/One Detectors, Counters, High Density Memory Elements. SRAM, DRAM, ROM, Serial access memories, Content Addressable Memory.

UNIT- V:

Programmable logic Devices: PLD's, CPLD's, FPGAs, Standard Cells, sea of gates, Design Approach, Parameters influencing low power design.

CMOS Testing: Need and importance of testing, CMOS Testing, Test Principles, Design Strategies for test, Chip level Test Techniques.

TEXTBOOKS:

1. Essentials of VLSI circuits and systems – Kamran Eshraghian, Eshraghian Dougles and A. Pucknell, PHI, tion.

2. CMOS VLSI Design – A Circuits and Systems Perspective, Neil.H.E.Weste a, David Harris, Ayan Banerjee, 3rd Ed, Pearson Education, 2009.

3 .VLSI Technology SM SZE.

REFERENCES:

1. Introduction to VLSI Systems: A Logic, Circuit and System Perspective - Ming-BO Lin, CRC Press, 2011

2. CMOS logic circuit Design - John .P. Uyemura, Springer, 2007.

3. Modern VLSI Design - Wayne Wolf, Pearson Education, 3rd Edition, 1997.

4. VLSI Design- K .Lal Kishore, V. S. V. Prabhakar, I.K International, 2009.

5. Introduction to VLSI - Mead & Convey, BS Publications, 2010.

Course Outcomes:

- Upon successfully completing the course, the student should be able to:
- Acquire qualitative knowledge about the fabrication process of integrated circuit using MOS transistors.
- Choose an appropriate inverter depending on specifications required for a circuit
- Draw the layout of any logic circuit which helps to understand and estimate parasitic of any logic circuit
- Design different types of logic gates using CMOS inverter and analyze their transfer characteristics
- Provide design concepts required to design building blocks of data path using gates.
- Design simple memories using MOS transistors and can understand
- Design of large memories.
- Design simple logic circuit using PLA, PAL, FPGA and CPLD.
- Understand different types of faults that can occur in a system and learn the concept of testing and adding extra hardware to improve testability of system

TEACHING NOTES

UNIT-I

INTRODUCTION

Contents:

- Introduction to IC Technology
- MOS, PMOS, NMOS, CMOS & Bi CMOS technologies
- Oxidation, Lithography,

- Diffusion, Ion implantation,
- Metallisation, Encapsulation,
- Probe testing
- Integrated Resistors and Capacitors.

BASIC ELECTRICAL PROPERTIES

- Basic Electrical Properties of MOS and Bi CMOS Circuits
- I_{ds}-V_{ds} relationships
- MOS transistor threshold Voltage
- g_m, g_{ds}, figure of merit
- Pass transistor
- NMOS Inverter
- Various pull ups
- CMOS Inverter analysis and design
- Bi-CMOS Inverters.

UNIT-I

Introduction to VLSI Technology:

In 1958, the first IC flipflop with two transistors was built by jack kilby at texas Instruments. In 2003, Intel Pentium IV processor contained 55 million transistors and a 512 mbit DRAM contained more than half a billion transistors. This corresponds to the compound annual growth of 53% over 45 years.

This incredible growth has come from the steady miniaturization of transistors and improvement in the manufacturing process. As transistors become smaller, they also become faster, dissipate less power and cheaper to manufacture. Bipolar transistors which were first developed are reliable, less noise, more power efficient. Early IC's used only bipolar transistors. Transistors can be viewed as electrically controlled switches with a control terminal and two other terminals that are connected or disconnected depending on the voltage applied to the control.

Bipolar transistors require a small current into the control(base) terminal to switch much larger currents between other two(emitter and collector) terminals. The quiescent power dissipated by these base current limits the maximum no of transistors that can be integrated onto a single die.

Metal oxide semiconductor field effect transistor (MOSFET's) have a very good advantage that they almost draw zero current while idle. Two flavors of MOS are NMOS and PMOS which use n type and p type dopants.

Frank wanlass at Fairchild described the first logic gates using MOSFET's in 1963, using nmos and pmos transistors, (cmos). This circuits used discrete transistors but consumed only nanowatts of power,(six orders or magnitude less than their bipolar counter parts)

MOS transistors are advantageous because

1. they occupies less area
2. simple fabrication process
3. low cost

Early process used only PMOS transistor but suffered form poor performance, yield and reliability. Processes using nmos transistors become dominant in 1970's. NMOS process was less expensive than cmos but nMOS logic gates still consumed power while idle.

Power consumption became a major issue in 1980's as hundreds of thousands of transistors were integrated onto a single die CMOS process were widely adopted and have essentially replaced nMOS and bipolar process for nearly all digital logic applications. Gorden Moore observed in 1965 that plotting the number of transistors that can be made economically fabricated on chip gives a straight line on a semilogarthmic scale. He found that transistor count doubling fot every 18 months. This observation is called Moore's law.

LEVELS OF INTEGRATION OF CHIPS:

SSI(small scale Integration)---------less than 10 gates

MSI(Medium scale Integration)----------upto 1000 gates

LSI(Large scale Integration)----------upto 10,000 gates

VLSI(very large scale Integration)----------More than 10,000 gates

ULSI(ultra large scale Integration)--------millions of gates

Pentium IV uses transistors with minimum dimensions of 130nm in 2003 and even further smaller dimensions now a days. This scaling cannot go on forever because transistors cannot be smallest than atoms. In the early 1990's experts agreed that scaling would continue for at least a decade.

PROCESS TECHNOLOGIES USED TODAY:

1. CMOS (complementary metal oxide semiconductor) technology.
2. Bipolar Technology
3. Bi CMOS Technology
4. SOI (Silicon on Insulator) Technology.

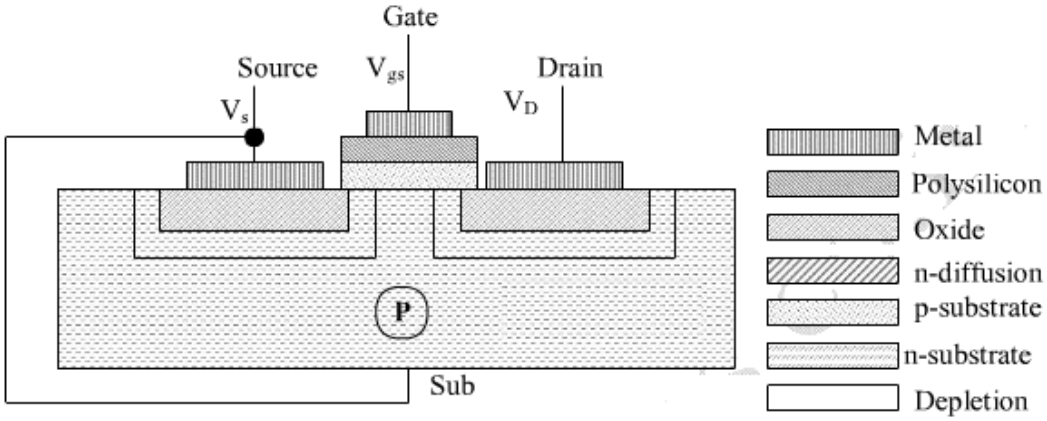
In metal Oxide semiconductor technology, metal gate electrode placed on top of an oxide insulator. In today's CMOS process, instead of metal, the gate electrode is comprise of a different material, polysilicon as it can withstand high processing temperatures.

The majority of IC's manufactured are cmos circuits due to three characteristics of cmos devices like high noise immunity, low static power, high density.

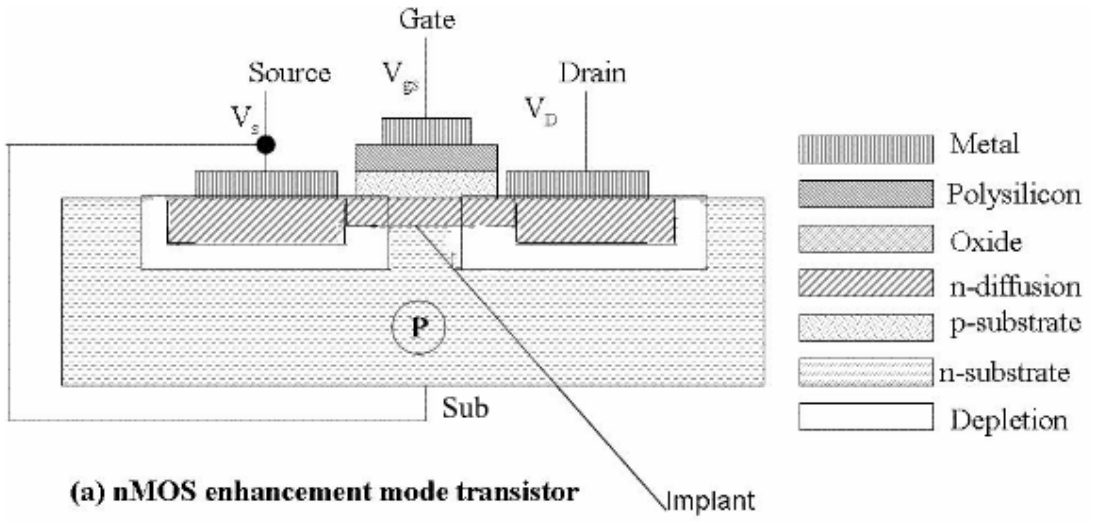
1.2) Operation of MOS,PMOS,NMOS transistors :

BASIC MOS TRANSISTORS

MOS means Metal oxide semiconductor technology. The basic n MOS enhancement and depletion mode transistors are shown in the figures below.



(a) nMOS enhancement mode



(a) nMOS enhancement mode transistor

n MOS devices are formed in a P-type substrate of moderate doping level. The source and drain regions are formed by diffusing n – type impurities through suitable marks into these areas to give the desired n – impurity concentration and give rise to depletion regions which extend mainly in the more lightly doped P

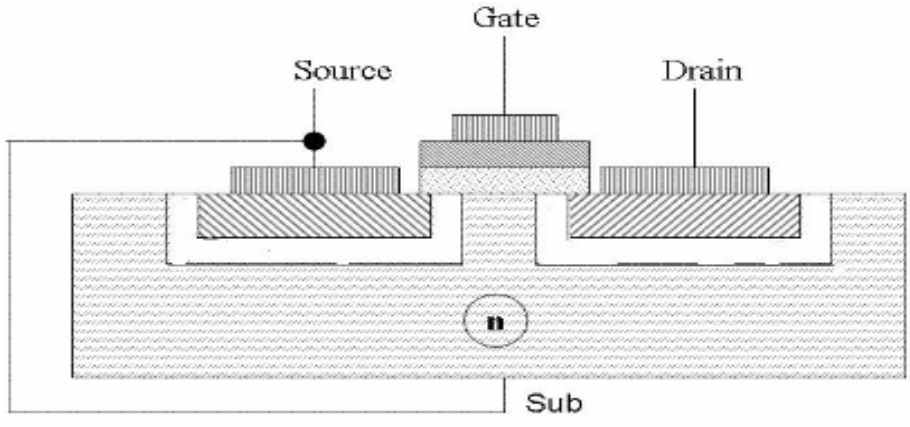
region as shown Thus, source and drain are isolated from one another by two diodes. Connections to the Source and drain are made by a deposited metal layer.

Consider the enhancement mode device first. A Polysilicon gate is deposited on a layer of insulation over the region between source and drain. The basic enhancement mode is shown in the figure (a) in which the channel is not established and the device is in a non-conducting condition,

$V_D = v_S = V_{ge} = 0$. If this gate is connected to suitable positive voltage with respect to the source, then the electric field established between the g and the substrate gives rise to a charge inversion region in the substrate under the gate insulation and a conducting path or channel is formed between source and drain.

The channel may also be established so that it is present under the condition $V_{gs} = O$ by implanting suitable impurities in the region between source and drain during manufacture. This is the n MOS depletion mode transistor. The source and drain are connected by a conducting channel, but the channel may now be closed by applying a suitable negative voltage to the gate.

In both cases, variations of the gate voltage allow control of any current flow between source and drain.



(c) pMOS enhancement mode transistor

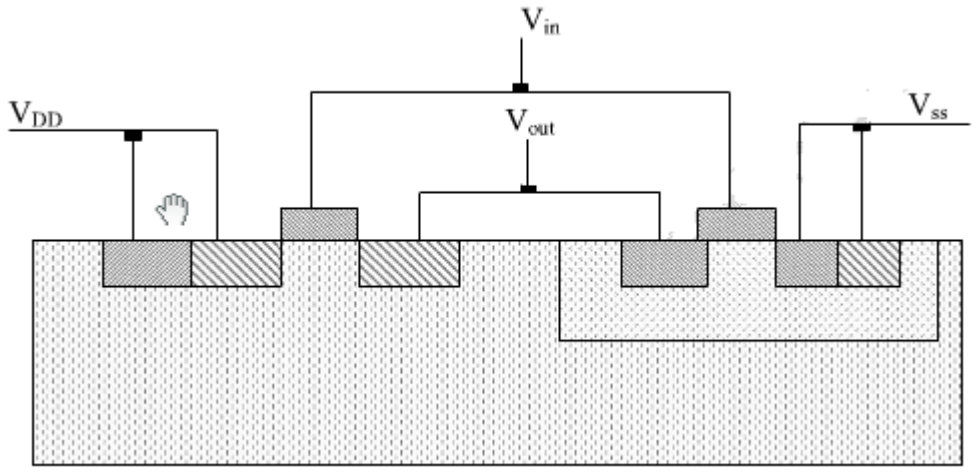
Figure 1.4 MOS transistors (V_D=0V. Source gate and substrate to 0 V)

The above figure shows the PMOS transistor structure for an enhancement mode device. In this case, the substrate is of n – type material and the source and drain diffusions are consequently P- type. In the figure, the conditions are for an unbiased device, however the application of negative voltage of suitable magnitude $(|V_t|)$ between gate and source will give rise to the formation of a channel (P-type) between the source and drain and current may then flow if the drain is made negative with respect to the source. In this case current is carried by the holes as opposed to electrons. PMOS transistors are slower than n MOS since whole mobility up is lees than electron mobility μ_n.

CMOS

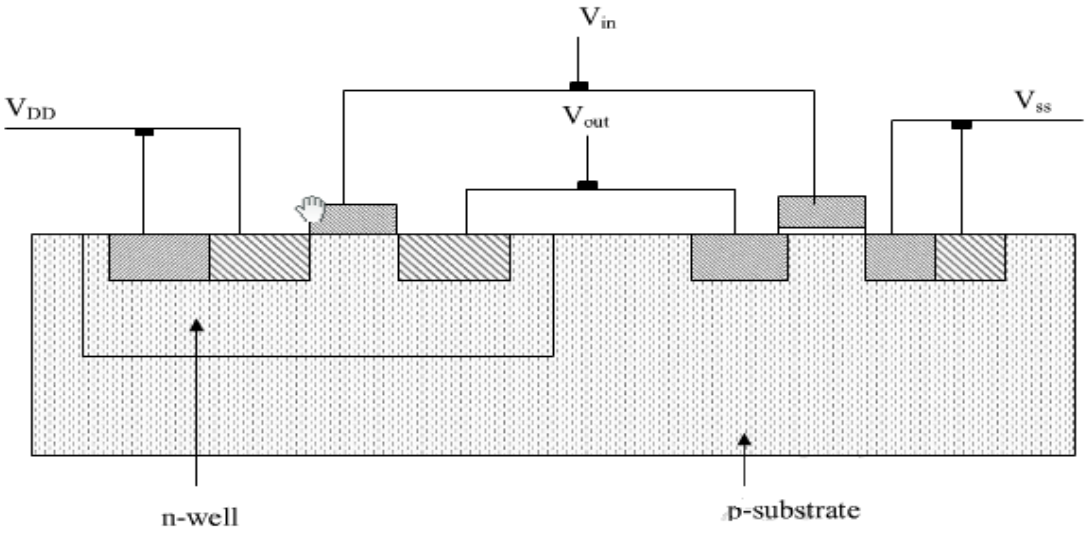
There are two types of fabrication of CMOS basically. P – well and n – well inverters are the major ones.

In CMOS P- well inverter, the structure consist of an n – type substrate in which P- devices may be formed by suitable marking and diffusion and, in order to accommodate n – type devices, a deep P – well is diffused into the n – type substrate.



CMOS p-well inverter showing V_{DD} and V_{SS} substrate

N – well CMOS circuits are superior to P-well because of the lower substrate bias effects on transistor threshold voltage and inherently lover Parasitic capacitances associated with source and drain regions. The CMOS n – well inverter is shown in figure below.



Cross-sectional view of n well CMOS

1.3) Fabrication of CMOS inverter

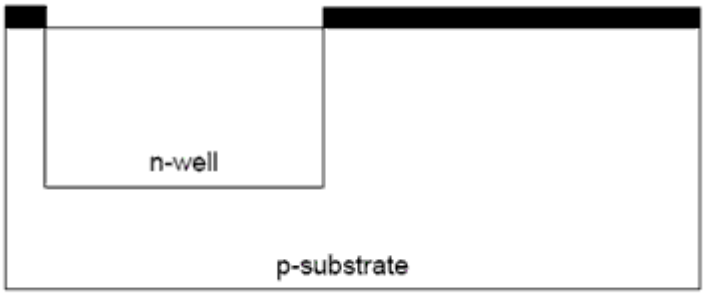
Basic CMOS Technology

In early days of technology, the control gate of the MOS transistor was made with aluminum instead of polycrystalline silicon. It was difficult to align the metal over the channel precisely; an offset in one direction or other would create a non-functioning of the transistor. To overcome these problems, the poly-silicon gate was introduced. This polysilicon would be deposited before source/drain diffusion. During the diffusion, source and drain regions are self-aligned with respect to the gate. This self-alignment structure reduces the device size. In addition, it eliminates the large overlap capacitance between gate and drain, while maintaining a continuous inversion layer between source and drain. In the case of metal gate process, Al deposition has to be carried out almost at the end of fabrication because further high temperature processing would melt Al. In case of self-aligned poly silicon gate technology, these restrictions are also circumvented.

Basic n-well CMOS process:

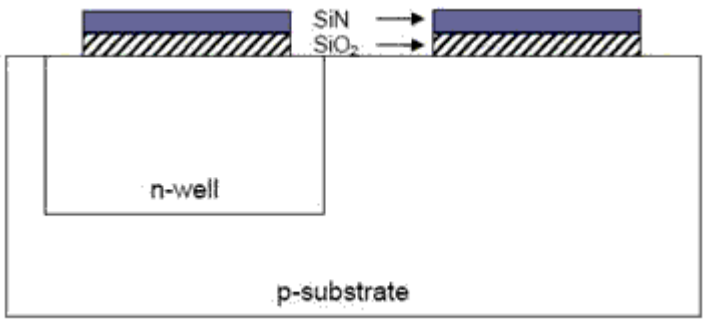
In a standard n-well process, one of the first things made is the n-well in a p type substrate. Once the n-well is created, the active areas can be defined. The MOSFET is build within this active area. A very thin layer of silicon dioxide is grown on the surface. This will be used to insulate the gate from the surface. The thin layer of SiO_2 is grown and covered with Si_3N_4. This will act as a mask during the subsequent channel stop implant and field oxide growth. The channel stop implant is to prevent conduction between unrelated transistor source/drains. A thick additional layer oxide grows in both directions vertically where Si_3N_4 is absent. Layer of silicon dioxide under the polysilicon gate (which will be created later) is known as gate oxide and that is not directly under the gate of a transistor is known as field oxide. The field oxide provides isolation between transistors. A threshold adjustment implant would be the next process step. This is carried out to balance off the threshold voltage differences. The P-MOS results in a higher threshold voltage level than nMOS with normal doping concentrations. With additional negative charges buried inside the channel, V_T for pMOS could be controlled.

(a)



Formation of n-well

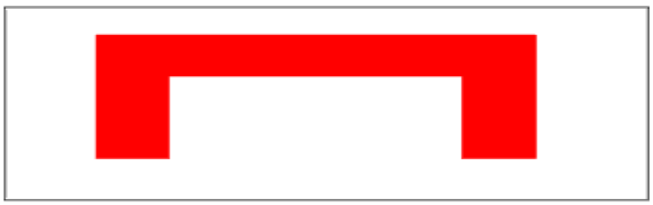
(b)



Gate oxide covered with silicon nitride in the active areas

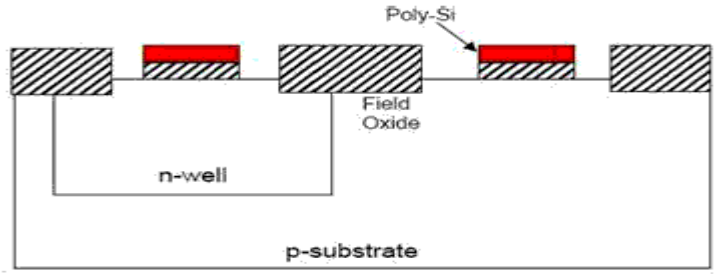
Polysilicon deposition is carried out and gate definition is then completed using the mask shown in fig (c). Note that the connection between two gate inputs in a CMOS inverter is achieved using the poly silicon. The source and drain diffusions for pMOS is carried out using p-type diffusion. Boron is the most popular element used for this step. Similarly, source and drain diffusions for nMOS is carried out using n-type diffusion. Phosphorous and Arsenic can both be used for this step. Additional oxide is created, and then the contact holes are cut in the oxide down to the diffusions and polysilicon. These contacts can be filled by metal permitted to flow into the holes. The drains of pMOS and nMOS transistors are connected by a metal line in order to take the output from the CMOS inverter.

©



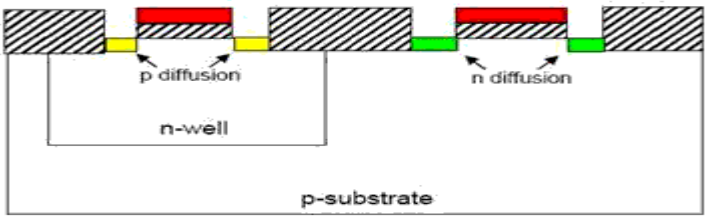
Top view of Poly silicon mask

(d)

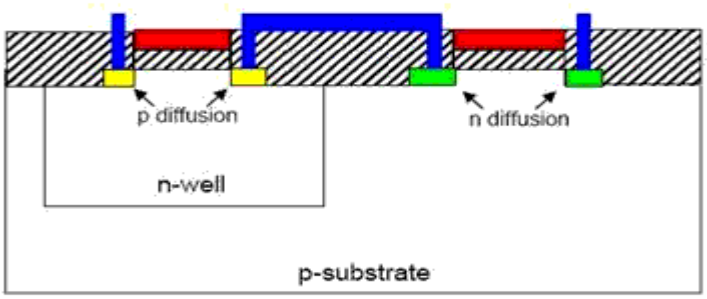


Poly silicon gate definition is completed

(e)



Transistor source/drain diffusion is completed



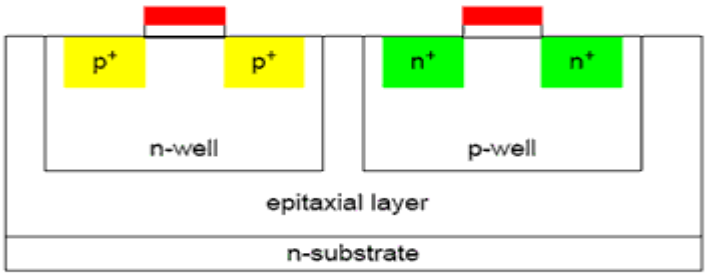
(F)**Cross section of a CMOS inverter in an n-well process**

P-well process

Prior to the n-well process p-well process was popular. P-well process is preferred in circumstances where balanced characteristics of the nMOS and pMOS are needed. It has been observed that the transistors in the native substrate tend to have better characteristics than that was made in a well. Because p devices inherently have lower gain than devices, n well process amplifies this difference while a p-well process moderates the difference. The standard p-well process steps are is similar to n-well process, except that a p-well is implanted instead of an n-well as a first step. Once the p-well is created, the active areas and subsequently poly gates can be defined. Later diffusions can be carried out to create source and drain regions. Finally, metal is deposited and patterned for contacts.

Twin-Tub process:

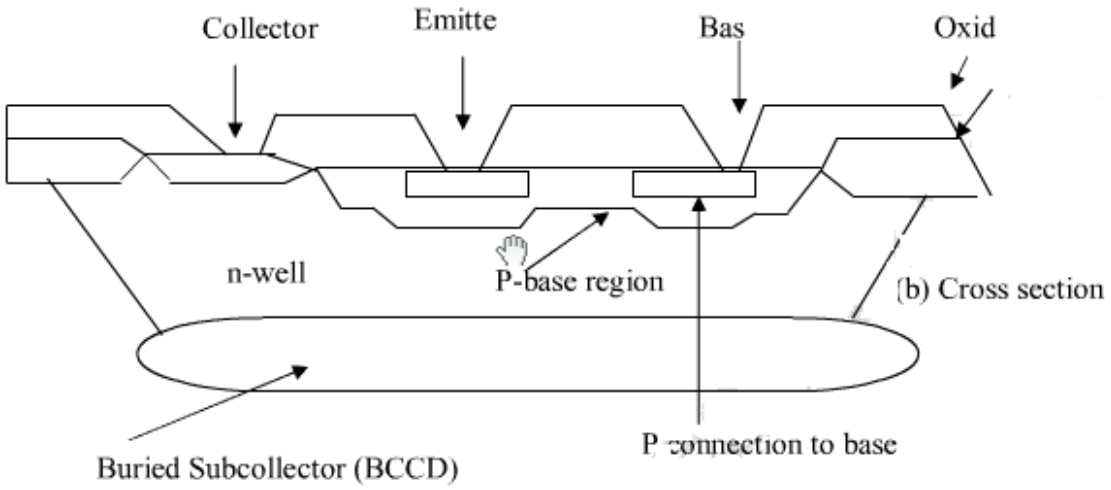
It is also possible to create both a p-well and an n-well for the n-MOSFET's and p-MOSFET respectively in the twin well or twin tub technology. Such a choice means that the process is independent of the dopant type of the starting substrate (provided it is only lightly doped).



A simplified sketch of twin-well CMOS process cross section

BICMOS TECHNOLOGY

The limited drive capabilities of MOS transistors can be overcome with the BiCMOS npn transistor consists of a P^+ base region, n^+ collector area and the buried sub collector (BCCD)



Arrangement of BICMOS npn transistor

Single crystal Si manufacture:

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 oat zone technique is used for producing specialty wafers that have low oxygen impurity concentration.

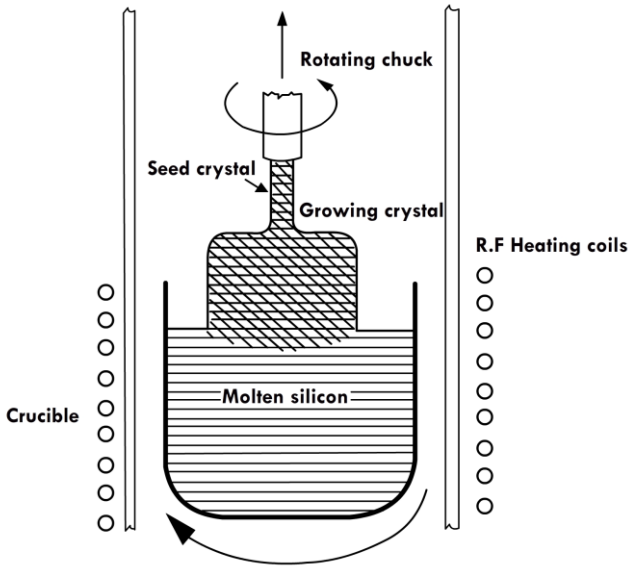
1.Czochralski crystal growth technique:

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 SiO2 or SiNx. The drawback is that at the high temperature the inner liner of the crucible also starts melting and has to

replace periodically. The Figure Schematic of the Czochralski growth technique. The polycrystalline silicon is melted and a single crystal seed is then used to nucleate a single crystal ingot. The seed crystal controls the orientation of the single crystal.

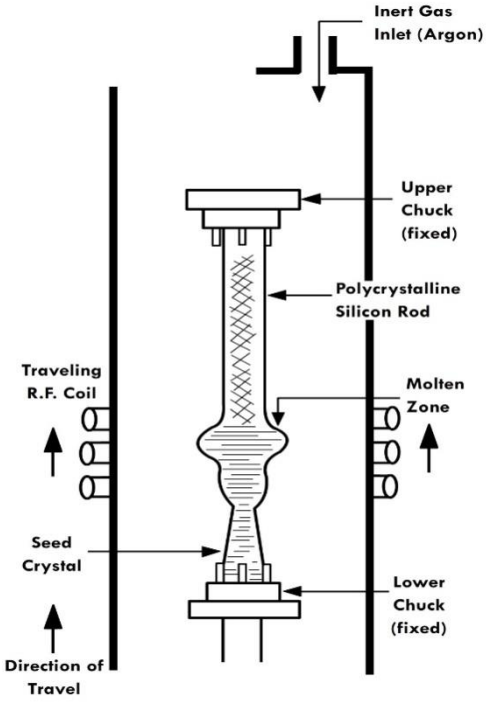


Above Figure Schematic of the Czochralski growth technique. The polycrystalline silicon is melted and a single crystal seed is then used to nucleate a single crystal ingot. The seed crystal controls the orientation of the single crystal.

Above Figure Single crystal Si ingot. This is further processed to get the wafers that are used for fabrication. Furnace is heated above 1500 ^{o}C, since Si melting point is 1412 ^{o}C. A small seed crystal, with the desired orientation of the final wafer, is dipped in the molten Si and slowly withdrawn by the crystal pulling mechanism. The seed crystal is also rotated while it is being pulled, to ensure uniformity across the surface. 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. The process control, i.e. speed of withdrawal and the speed of rotation of the crystal puller, is crucial to obtain a good quality single crystal. There is a feedback system that control this process. Similarly there is another ambient gas control system. The final solidified Si obtained is the single crystal ingot. A 450 mm wafer ingot can be as heavy as 800 kg. A picture of an ingot is show in above figure.

2. 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 figure6. 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.



The above Figure Schematic of the float zone technique. The polycrystalline ingot is fused with a seed crystal and locally melted by a traveling radio frequency coil. As the ingot melts and resolidifes it has the same orientation as the seed.

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 manufacturing

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. There are two types of flats.

1. **Primary flat** - this is ground relative to a specific crystal direction. This acts as a visual reference to the orientation of the wafer.
2. **Secondary flat** - this used for identification of the wafer, dopant type and orientation.

The different flat locations are shown in figure 7. *p*-type (111) Si has only one flat (primary flat) while all other wafer types have two flats (with different orientations of the secondary flats). The primary flat is typically longer than the secondary flat. Consider some typical specs of 150 *mm* wafers, shown in table 4. Bow refers to the flatness of the wafer while Δt refers to the thickness variation across the wafer.

After making the flats, the individual wafers are sliced per the required thickness. *Inner diameter (ID) slicing* is the most commonly used technique. The cutting edge is located on the inside of the blade, as seen in figure 8. Larger wafers are usually thicker, for mechanical integrity.

After cutting, the wafers are chemically etched to remove any damaged and

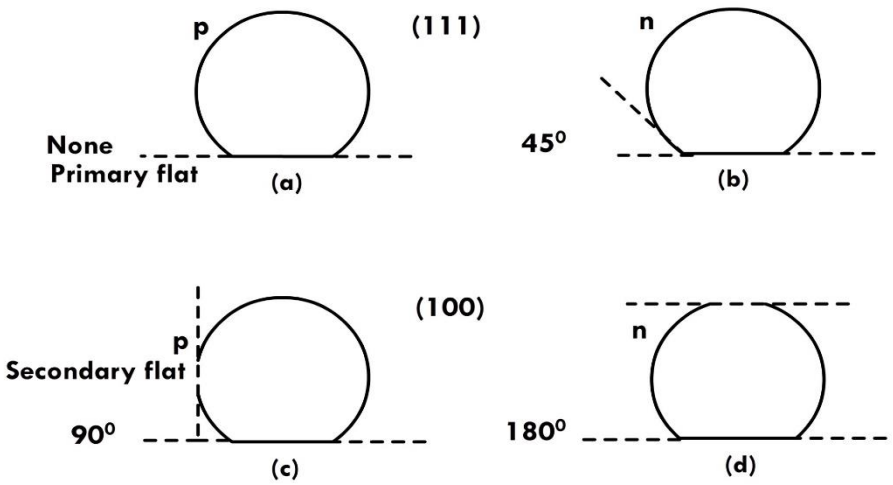
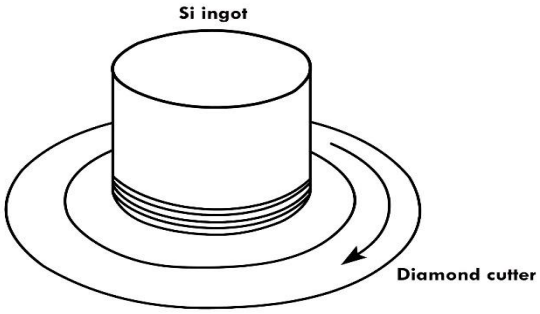


Figure 7: Flats for the different wafer types and orientations. All orientations and doping types have a primary flat, while there are different secondary flats for different types (a) p(111) (b) n(111) (c) p(100) and (d) n(100). Adapted from *Microchip fabrication - Peter van Zant.*

Table 4: Specs of a typical 150 mm wafer

Specs	**Value**
Diameter	150 ± 0.5 mm
Thickness	675 ± 25 μm
Orientation	100 ± 1°
Bow	60 μm
Δt	50 μm
Primary flat	55-60 *mm*
Secondary flat	35-40 *mm*



The above figure Inner diameter wafer slicing, used for cutting the ingots into individual wafers. The thickness is slightly higher than the final required thickness to account for material loss due to polishing. Adapted from *Microchip fabrication - Peter van Zant.*
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 SiO_2 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.

Poly Si manufacture

The starting material for Si wafer manufacture is called *Electronic grade Si* (EGS). This is an ingot of Si that can be shaped and cut into the final wafers. EGS should have impurity levels of the order of *ppb*, with the desired doping levels, so that it matches the chemical composition of the final Si wafers. The doping levels are usually back calculated from resistivity measurements. To get EGS, the starting material is called *Metallurgical grade Si* (MGS). The first step is the synthesis of MGS from the ore.

The starting material for Si manufacture is *quartzite* (SiO_2) or *sand*. The ore is reduced to Si by mixing with coke and heating in a submerged electrode arc furnace. The SiO_2 reacts with excess C to first form SiC. At high temperature, the SiC reduces SiO_2 to form Si. The overall reaction is given by

$$SiC\,(s) + SiO_2\,(s) \rightarrow Si\,(l) + SiO\,(g) + CO\,(g) \qquad (1)$$

The *Si*(*l*) formed is removed from the bottom of the furnace. This is the MGS and is around 98% pure. The schematic of the reducing process is shown in figure 1. Typical impurities and their concentrations in MGS is tabulated in
2. MGS is used for making alloys. From table 2 it can be seen that the main

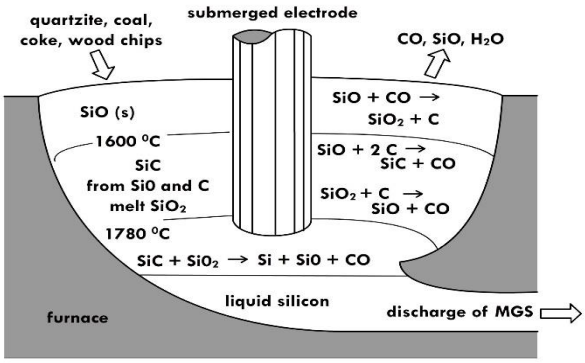


Figure above: Schematic of the submerged arc electrode process. SiO_2 is mixed with coke and heated. It first forms SiC, which further reacts with the remaining SiO_2 forming silicon. The temperature is maintained above the melting point of silicon so that the molten semiconductor is removed from the bottom. Adapted from *Synthesis and purification of bulk semiconductors* -Barron and Smith

Metallic impurities are Al and Fe. Further purification is needed to make EGS since the impurity concentration must be reduced to *ppb* levels.

One of the techniques for converting MGS to EGS is called the **Seimens process**. In this the Si is reacted with HCl gas to form tricholorosilane, which is in gaseous form.

$$Si\,(s) + 3HCl\,(g) \rightarrow SiHCl_3\,(g) + H_2\,(g) \qquad (2)$$

This process is carried out in a *fluidized bed reactor* at 300˚C, where the trichlorosilane gas is removed and then reduced using H_2 gas.

$$2SiHCl_3\,(g) + 2H_2\,(g) \rightarrow 2Si\,(s) + 6HCl\,(g) \qquad (3)=$$

The process flow is shown in figure 2. A Si rod is used to nucleate the reduced Si obtained from the silane gas, as shown in figure 3. During the conversion of silicon to trichlorosilane impurities are removed and process can be cycled to increase purity of the formed Si. The final material obtained is the EGS. This is a polycrystalline form of Si, like MGS, but has much smaller impurity levels, closer to what is desired in the final single crystal wafer. The impurities in EGS are tabulated in 3. EGS is still polycrystalline and needs to be converted into a single crystal Si ingot for producing the wafers.

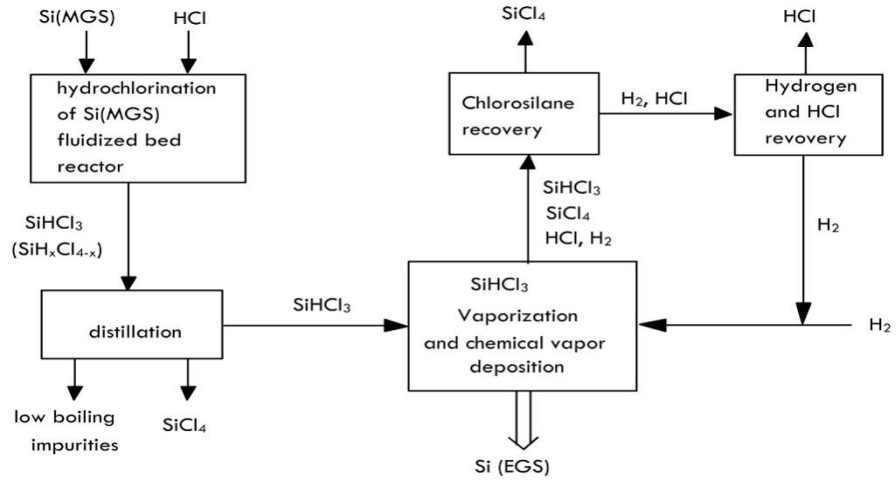


Figure 2: Schematic of the process to purify MGS to obtain EGS. The process involves conversion of silicon to trichlorosilane gas, which is purified, and then reduced to obtain silicon. Adapted from *Synthesis and purification of bulk semiconductors - Barron and Smith*

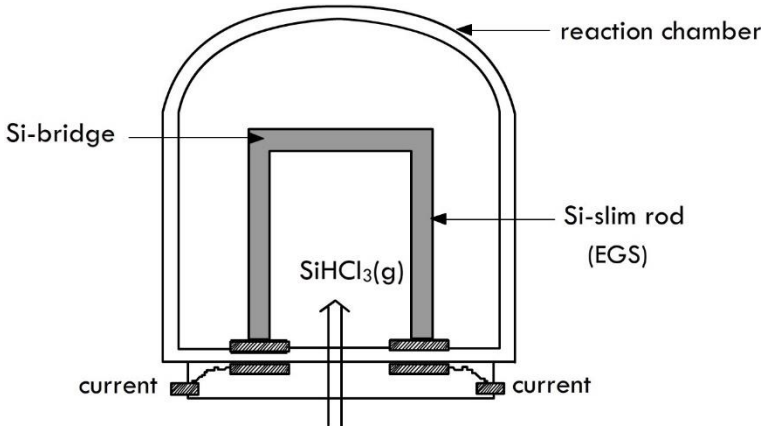


Figure 3: The Seimens deposition reactor where the purified Si is condensed. This is the electronic grade Si, same purity level as Si wafers, but polycrystalline. Adapted from *Synthesis and purification of bulk semiconductors -*

Barron and Smith

Table 3: Impurities in EGS, after purification from MGS. Compared to table 2, the concentration levels of the metals have dropped to *ppb* levels.

Element	**Concentration (*ppb*)**
As	<0.001
Sb	<0.001
B	<0.1
C	100-1000
Cu	0.1
Fe	0.1-1
O	100-400
P	<0.3

IC Fabrication Process Steps

The fabrication of integrated circuits consists basically of the following process steps:

- **Lithography:** The process for pattern definition by applying thin uniform layer of viscous liquid (photo-resist) on the wafer surface. The photo-resist is hardened by baking and than selectively removed by projection of light through a reticle containing mask information.
- **Etching:** Selectively removing unwanted material from the surface of the wafer. The pattern of the photo-resist is transferred to the wafer by means of etching agents.
- **Deposition:** Films of the various materials are applied on the wafer. For this purpose mostly two kind of processes are used, physical vapor deposition (PVD) and chemical vapor deposition (CVD).
- **Chemical Mechanical Polishing:** A planarization technique by applying a chemical slurry with etchant agents to the wafer surface.

- **Oxidation:** In the oxidation process oxygen (dry oxidation) or H^2O (wet oxidation) molecules convert silicon layers on top of the wafer to silicon dioxide.
- **Ion Implantation:** Most widely used technique to introduce dopant impurities into semiconductor. The ionized particles are accelerated through an electrical field and targeted at the semiconductor wafer.
- **Diffusion:** A diffusion step following ion implantation is used to anneal bombardment-induced lattice defects.
- **Metallization:** Metallization is the final step in the wafer processing sequence. Metallization is the process by which the components of IC's are interconnected by aluminium conductor. This process produces a thin-film metal layer that will serve as the required conductor pattern for the interconnection of the various components on the chip. Another use of metallization is to produce metalized areas called bonding pads around the periphery of the chip to produce metalized areas for the bonding of wire leads from the package to the chip.

Silicon Dioxide

Dry Oxidation

During dry oxidation, the wafer is placed in a pure oxygen gas (O^2) environment and the chemical reaction which ensues is between the solid silicon atoms (Si) on the surface of the wafer and the approaching oxide gas

$$Si + O_2 \rightarrow SiO_2. \tag{29}$$

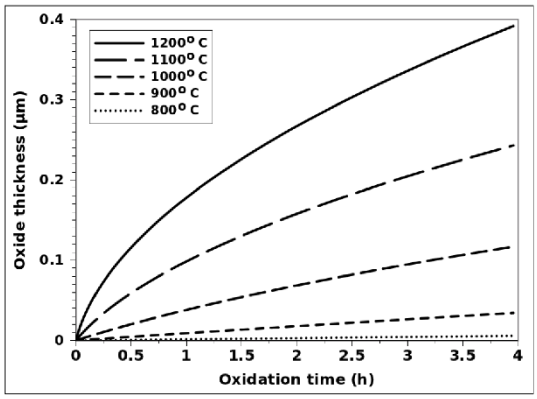


Figure : Oxide thickness versus oxidation time for dry (O^2) oxidation of a (100) oriented silicon wafer under various temperatures.

Above Figure shows the oxide thickness as a function of oxidation time for dry oxidation. It can be noted that the oxidation rate does not exceed $\sim$150nm$/$h, making it a relatively slow process which can be accurately controled in order to achieve a desired thickness. 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.

2.2.1.2 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^2) released as a byproduct

$$Si + 2H_2O \rightarrow SiO_2 + 2H_{2(g)}. \quad (30)$$

Figure shows the oxide thickness as a function of oxidation time for wet oxidation processing.

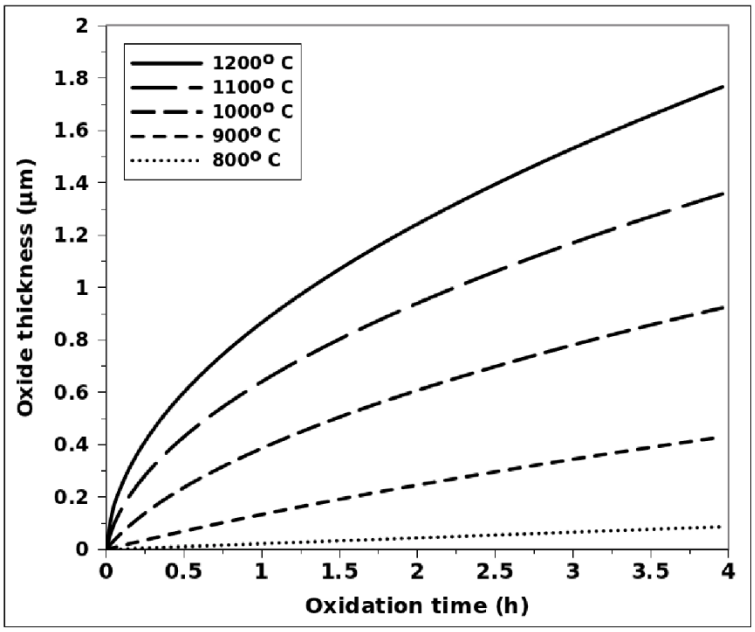


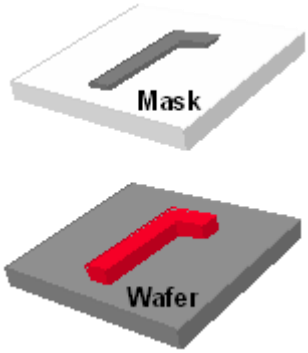
Figure Oxide thickness versus oxidation time for wet (H^2O) oxidation of a (100) oriented silicon wafer under various temperatures.

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.

Lithography

The word lithography comes from the Greek lithos, meaning stones, and graphia, meaning to write. It means quite literally writing on stones. In the case of semiconductor lithography (also called photolithography) our stones are silicon wafers and our patterns are written with a light sensitive polymer called a photoresist. To build the complex structures that make up a transistor and the many wires that connect the millions of transistors of a circuit, lithography

and etch pattern transfer steps are repeated at least 10 times, but more typically are done 20 to 30 times to make one circuit. Each pattern being printed on the wafer is aligned to the previously formed patterns and slowly the conductors, insulators, and selectively doped regions are built up to form the final device.



The general sequence of processing steps for a typical photolithography process is as follows: substrate preparation, photoresist spin coat, prebake, exposure, post-exposure bake, development, and postbake. A resist strip is the final operation in the lithographic process, after the resist pattern has been transferred into the underlying layer. This sequence is shown diagrammatically in below Figure

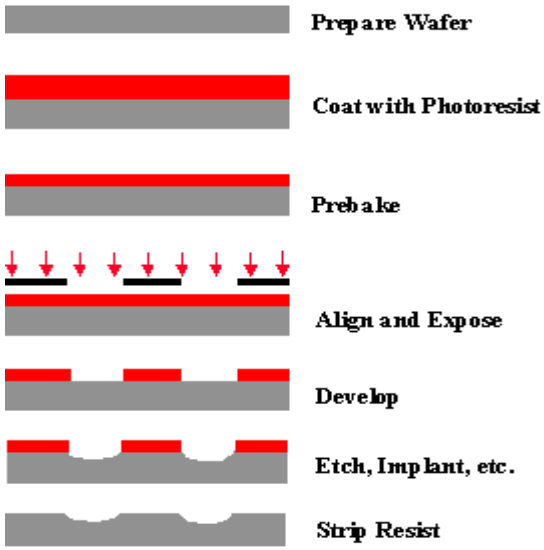


Figure Example of a typical sequence of lithographic processing steps (with no post-exposure bake in this case), illustrated for a positive resist.

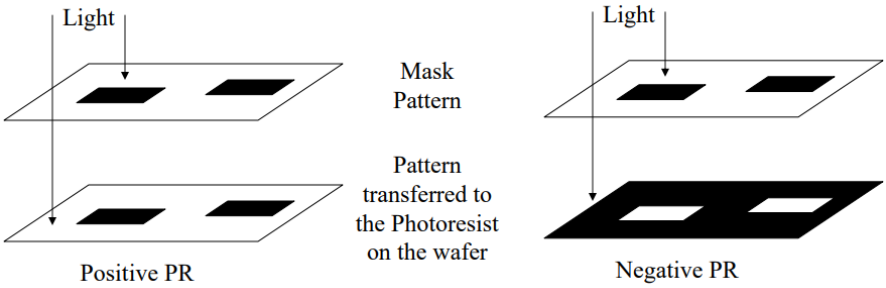
Lithography and Photoresists

Used for Pattern transfer into oxides, metals, semiconductors. 3 types of Photoresists (PR):

1.) Positive: PR pattern is same as mask. On exposure to light, light degrades the polymers (described in more detail later) resulting in the photoresist being more soluble in developers. The PR can be removed in inexpensive solvents such as acetone.

2.) Negative: PR pattern is the inverse of the mask. On exposure to light, light polymerizes the rubbers in the photoresist to strengthen it's resistance to dissolution in the developer. The resist has to be removed in special stripping chemicals. These resists tend to be extremely moisture sensitive.

3.) Combination: Same photoresist can be used for both negative and positive pattern transfer. Can be removed in inexpensive solvents.



Etching

Etching is used to remove material selectively in order to create patterns. The pattern is defined by the etching mask, because the parts of the material, which should remain, are protected by the mask. The unmasked material can be removed either by wet (chemical) or dry (physical) etching. Wet etching is strongly isotropic which limits its application and the etching time can be controlled difficultly. Because of the so-called under-etch effect, wet etching is not suited to transfer patterns with sub-micron feature size. However, wet etching has a high selectivity (the etch rate strongly depends on the material) and it does not damage the material. On the other side dry etching is highly anisotropic but less selective. But it is more capable for transfering small structures.

Deposition

Diffusion of Dopant Impurities

The process of junction formation, that is transition from p to n type or vice versa, is typically accomplished by the process of diffusing the appropriate dopant impurities in a high temperature furnace. Impurity atoms are introduced onto the surface of a silicon wafer and diffuse into the lattice because of their tendency to move from regions of high to low concentration. Diffusion of impurity atoms into silicon crystal takes place only at elevated temperature, typically 900 to 1100°C.

Although these are rather high temperatures, they are still well below the melting point of silicon, which is at 1420°C. The rate at which the various impurities diffuse into silicon will be of the order of 1 micro meter per hour at a temperature range stated above, and the penetration depth that are involved in most diffusion processes will be of the order of 0.3 to 30 micro meter. At room temperature the diffusion process will be so extremely slow such that the impurities can be considered to be essentially frozen in place.

A method of p-n junction formation which was popular in the early days is the grown junction technique. In this method the dopant is abruptly changed in the melt during the process of crystal growth. A convenient technique for making p-n junction is the alloying of a

metal containing doping atoms on a semiconductor with the opposite type of dopant. This is called the alloyed junction technique. The p-n junction using epitaxial growth is widely used in ICs. An epitaxial grown junction is a sharp junction. In terms of volume of production, the most common technique for forming p-n junctions is the impurity diffusion process. This produces diffused junction. Along with diffusion process the use of selective masking to control junction geometry, makes possible the wide variety of devices available in the form of IC's. Selective diffusion is an important technique in its controllability, accuracy and versatility.

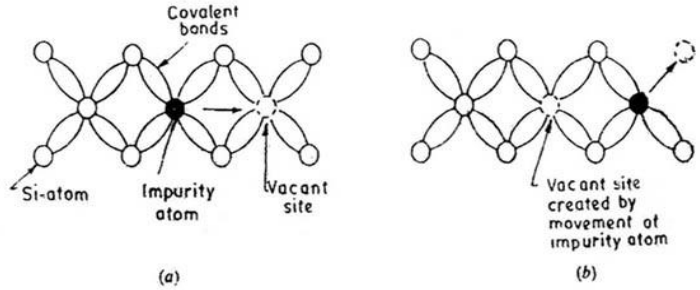
Nature of Impurity Diffusion

The diffusion of impurities into a solid is basically the same type of process as occurs when excess carriers are created non-uniformly in a semiconductor which cause carrier gradient. In each case, the diffusion is a result of random motion, and particles diffuse in the direction of decreasing concentration gradient The random motion of impurity atoms in a solid is, of course, rather limited unless the temperature is high. Thus diffusion of doping impurities into silicon is accomplished at high temperature as stated above.

There are mainly two types of physical mechanisms by which the impurities can diffuse into the lattice. They are

1. Substitutional Diffusion

At high temperature many atoms in the semiconductor move out of their lattice site, leaving vacancies into which impurity atoms can move. The impurities, thus, diffuse by this type of vacancy motion and occupy lattice position in the crystal after it is cooled. Thus, substitutional diffusion takes place by replacing the silicon atoms of parent crystal by impurity atom. In other words, impurity atoms diffuse by moving from a lattice site to a neighbouring one by substituting for a silicon atom which has vacated a usually occupied site as shown in the figure below.



Substitutional Diffusion By Dopant Impurities

Substitutional diffusion mechanism is applicable to the most common diffusants, such as boron, phosphorus, and arsenic. These dopants atoms are too big to fit into the interstices or voids, so the only way they can enter the silicon crystal is to substitute for a Si atom.

In order for such an impurity atom to move to a neighbouring vacant site, it has to overcome energy barrier which is due to the breaking of covalent bonds. The probability of its having enough thermal energy to do this is proportional to an exponential function of temperature. Also, whether it is able to move is also dependent on the availability of a vacant

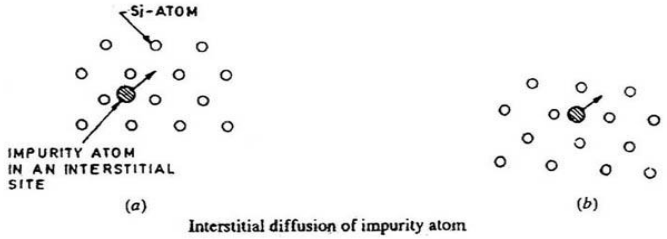
neighbouring site and since an adjacent site is vacated by a Si atom due to thermal fluctuation of the lattice, the probability of such an event is again an exponent of temperature.

The jump rate of impurity atoms at ordinary temperatures is very slow, for example about 1 jump per 10^{50} years at room temperature! However, the diffusion rate can be speeded up by an increase in temperature. At a temperature of the order 1000 degree Celsius, substitutional diffusion of impurities is practically realized in sensible time scales.

2. Interstitial Diffusion

In such, diffusion type, the impurity atom does not replace the silicon atom, but instead moves into the interstitial voids in the lattice. The main types of impurities diffusing by such mechanism are Gold, copper, and nickel. Gold, particularly, is introduced into silicon to reduce carrier life time and hence useful to increase speed at digital IC's.

Because of the large size of such metal atoms, they do not usually substitute in the silicon lattice. To understand interstitial diffusion, let us consider a unit cell of the diamond lattice of the silicon which has five interstitial voids. Each of the voids is big enough to contain an impurity atom. An impurity atom located in one such void can move to a neighbouring void, as shown in the figure below.



Interstitial diffusion of impurity atom

Interstitial Diffusion of Impurity Atom

In doing so it again has to surmount a potential barrier due to the lattice, this time, most neighbouring interstitial sites are vacant so the frequency of movement is reduced. Again, the diffusion rate due to this process is very slow at room temperature but becomes practically acceptable at normal operating temperature of around 1000 degree Celsius. It will be noticed that the diffusion rate due to interstitial movement is much greater than for substitutional movement. This is possible because interstitial diffusants can fit in the voids between silicon atoms. For example, lithium acts as a donor impurity in silicon, it is not normally used because it will still move around even at temperatures near room temperature, and thus will not be frozen in place. This is true of most other interstitial diffusions, so long-term device stability cannot be assured with this type of impurity.

Fick's Laws of Diffusion

The diffusion rate of impurities into semiconductor lattice depends on the following

- Mechanism of diffusion
- Temperature
- Physical properties of impurity
- The properties of the lattice environment
- The concentration gradient of impurities
- The geometry of the parent semiconductor

The behaviour of diffusion particles is governed by Fick's Law, which when solved for appropriate boundary conditions, gives rise to various dopant distributions, called profiles which are approximated during actual diffusion processes.

In 1855, Fick drew analogy between material transfer in a solution and heat transfer by conduction. Fick assumed that in a dilute liquid or gaseous solution, in the absence of convection, the transfer of solute atoms per unit area in a one-dimensional flow can be described by the following equation

$$\mathbf{F = -D\ \partial N(x,t)/\partial x = -\partial F(x,t)/\partial x}$$

where F is the rate of transfer of solute atoms per unit area of the diffusion flux density (atoms/cm^2-sec). N is the concentration of solute atoms (number of atoms per unit volume/cm^3), and x is the direction of solute flow. (Here N is assumed to be a function of x and t only), t is the diffusion time, and D is the diffusion constant (also referred to as diffusion coefficient or diffusivity) and has units of cm^2/sec.

The above equation is called Fick's First law of diffusion and states that the local rate of transfer (local diffusion rate) of solute per unit area per unit time is proportional to the concentration gradient of the solute, and defines the proportionality constant as the diffusion constant of the solute. The negative sign appears due to opposite direction of matter flow and concentration gradient. That is, the matter flows in the direction of decreasing solute concentration.

Fick's first law is applicable to dopant impurities used in silicon. In general the dopant impurities are not charged, nor do they move in an electric field, so the usual drift mobility term (as applied to electrons and holes under the influence of electric field) associated with the above equation can be omitted. In this equation N is in general function of x, y, z and t.

The change of solute concentration with time must be the same as the local decrease of the diffusion flux, in the absence of a source or a sink. This follows from the law of conservation of matter. Therefore we can write down the following equation

$$\mathbf{\partial N(x,t)/\partial t = -\partial F(x,t)/\partial x}$$

Substituting the above equation to 'F'. We get

$$\mathbf{\partial N(x,t)/\partial t = \partial/\partial x[D*\partial N(x,t)/\partial x]}$$

When the concentration of the solute is low, the diffusion constant at a given temperature can be considered as a constant.

Thus the equation becomes,

$$\mathbf{\partial N(x,t)/\partial t = D[\partial^2 N(x,t)/\partial x^2]}$$

This is Ficks second law of distribution.

Diffusion Profiles

Depending on boundary equations the Ficks Law has two types of solutions. These solutions provide two types of impurity distribution namely constant source distribution following complimentary error function (erfc) and limited source distribution following Gaussian distribution function.

Constant Source (erfc) Distribution

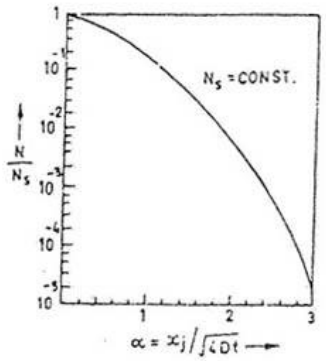
In this impurity distribution, the impurity concentration at the semiconductor surface is maintained at a constant level throughout the diffusion cycle. That is,

$$\mathbf{N\ (o,t) = N_S = Constant}$$

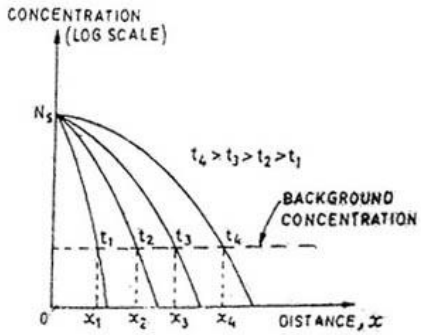
The solution to the diffusion equation which is applicable in this situation is most easily obtained by first considering diffusion inside a material in which the initial concentration changes in same plane as x=0, from N_S to 0. Thus the equation can be written as

$$\mathrm{N}\ \mathbf{(o,t) = N_S = Constant\ and\ N(x,t) = 0}$$

Shown below is a graph of the complementary error function for a range of values of its argument. The change in concentration of impurities with time, as described by the equation is also shown in the figure below. The surface concentration is always held at N_S, falling to some lower value away from the surface. If a sufficiently long time is allowed to elapse, it is possible for the entire slice to acquire a dopant level of N_S per m^3.



Complementary error function distribution.



Constant source diffusion profiles for various periods of time

Complimentary Error Function

If the diffused impurity type is different from the resistivity type of the substrate material, a junction is formed at the points where the diffused impurity concentration is equal to the background concentration already present in the substrate.

In the fabrication of monolithic IC's, constant source diffusion is commonly used for the isolation and the emitter diffusion because it maintains a high surface concentration by a continuous introduction of dopant.

There is an upper limit to the concentration of any impurity that can be accommodated at the semiconductor wafer at some temperature. This maximum

concentration which determines the surface concentration in constant source diffusion is called the solid solubility of the impurity.

Limited Source Diffusion or Gaussian Diffusion

Here a predetermined amount of impurity is introduced into the crystal unlike constant source diffusion. The diffusion takes place in two steps.

1. Predeposition Step – In this step a fixed number of impurity atoms are deposited on the silicon wafer during s short time.

2. Drive-in step – Here the impurity source is turned off and the amounts of impurities already deposited during the first step are allowed to diffuse into silicon water.

The essential difference between the two types of diffusion techniques is that the surface concentration is held constant for error function diffusion. It decays with time for the Gaussian type owing to a fixed available doping concentration Q. For the case of modelling the depletion layer of a p-n junction, the erfc is modelled as a step junction and the Gaussian as a linear graded junction. In the case of the erfc, the surface concentration is constant, typically the maximum solute concentration at that temperature or solid solubility limit.

Parameters which affect diffusion profile:

- **Solid Solubility** – In deciding which of the availability impurities can be used, it is essential to know if the number of atoms per unit volume required by the specific profile is less than the diffusant solid solubility.
- **Diffusion temperature** – Higher temperatures give more thermal energy and thus higher velocities, to the diffused impurities. It is found that the diffusion coefficient critically depends upon temperature. Therefore, the temperature profile of diffusion furnace must have higher tolerance of temperature variation over its entire area.
- **Diffusion time** – Increases of diffusion time, t, or diffusion coefficient D have similar effects on junction depth as can be seen from the equations of limited and constant source diffusions. For Gaussian distribution, the net concentration will decrease due to impurity compensation, and can approach zero with increasing diffusion tunes. For constant source diffusion, the net Impurity concentration on the diffused side of the p-n junction shows a steady increase with time.
- **Surface cleanliness and defects in silicon crystal –** The silicon surface must be prevented against contaminants during diffusion which may interfere seriously with the uniformity of the diffusion profile. The crystal defects such as dislocation or stacking faults

may produce localized impurity concentration. This results in the degradation of junction characteristics. Hence silicon crystal must be highly perfect.

Basic Properties of the Diffusion Process

Following properties could be considered for designing and laying out ICs.

- When calculating the total effective diffusion time for given impurity profile, one must consider the effects of subsequent diffusion cycles.
- The erfc and Gaussian functions show that the diffusion profiles are functions of ($x/\sqrt{Dt}$). Hence, for a given surface and background concentration, the junction depth x1 and x2 associated with the two separate diffusions having different times and temperature
- **Lateral Diffusion Effects** – The diffusions proceed sideways from a diffusion window as well as downward. In both types of distribution function, the side diffusion is about 75 to 80 per cent of the vertical diffusion.

Dopants and their Characteristics

The dopants selection affects IC characteristics. Boron and phosphorus are the basic dopants of most ICs. Arsenic and antimony, which are highly soluble in silicon and diffuse slowly, are used before epitaxial processing or as a second diffusion. Gold and silver diffuse rapidly. They act as recombination centres and thus reduce carrier life time.

Boron is almost an exclusive choice as an acceptor impurity in silicon since other p-type impurities have limitations as follows :

Gallium has relatively large diffusion coefficient in $Si0_2$, and the usual oxide window-opening technique for locating diffusion would be inoperative, Indium is of little interest because of its high acceptor level of 0.16 eV, compared with 0.01 eV for boron, which indicates that not all such acceptors would be ionized at room temperature to produce a hole. Aluminium reacts strongly with any oxygen that is present in the silicon lattice.

The choice of a particular n-type dopant is not so limited as for p-type materials. The n-type impurities, such as phosphorus, antimony and arsenic, can be used at different stages of IC processing. The diffusion constant of phosphorus is much greater than for Sb and As,

being comparable to that for boron, which leads to economies resulting from shorter diffusion times.

Dopants in VLSI Technology

The common dopants in VLSI circuit fabrication are boron, phosphorus. and arsenic. Phosphorus is useful not only as an emitter and base dopant, but also far gettering fast-diffusing metallic contaminants, such as Cu and An, which cause junction leakage current problems. Thus, phosphorus is indispensable in VLSI technology. However, n-p-n transistors made with arsenic-diffused emitters have better low-current gain characteristics and better control of narrow base widths than those made with phosphorus-diffused emitters. Therefore, in V LSI, the use of phosphorus as an active dopant in small, shallow junctions and low-temperature processing will be limited to its use as the base dopant of p-n-p device and as a gettering agent. Arsenic is the most frequently used dopant for the source and drain regions in n-channel MOSFETs.

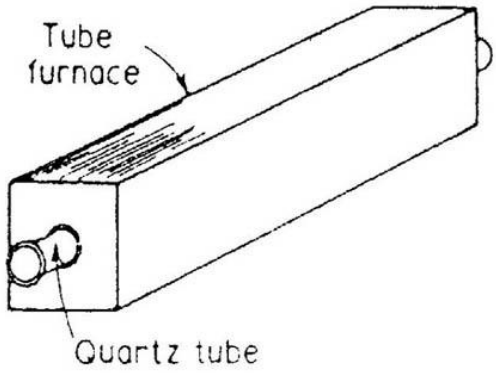
Diffusion Systems

Impurities are diffused from their compound sources as mentioned above. The method impurity delivery to wafer is determined by the nature of impurity source; Two-step diffusion is widely technique. Using this technique, the impurity concentration and profiles can be carefully controlled. The type of impurity distribution (erfc or Gaussian) is determined by the choice of operating conditions.

The two-step diffusion consists of a deposition step and a drive-in step. In the former step a constant source diffusion is carried out for a short time, usually at a relatively low temperatures, say, 1000°C. In the latter step, the impurity supply is shutoff and the existing dopant is allowed to diffuse into the body of the semiconductor, which is now held at a different temperature, say 1200°C, in an oxidizing atmosphere. The oxide layer which forms on tire surface of the wafer during this step prevents further impurities from entering, or those already deposited, from diffusing out. The final impurity profile is a function of diffusion condition, such as temperature, time, and diffusion coefficients, for each step.

- **Diffusion Furnace**

For the various types of diffusion (and also oxidation) processes a resistance-heated tube furnace is usually used. A tube furnace has a long (about 2 to 3 meters) hollow opening into which a quartz tube about 100,150 mm in diameter is placed as shown in the figure below.



Diffusion Furnace

The temperature of the furnace is kept about1000°C. The temperature within the quartz furnace tube can be controlled very accurately such that a temperature within 1/2°C of the set-point temperature can be maintained uniformly over a "hot zone" about 1 m in length. This is achieved by three individually controlled adjacent resistance elements. The silicon wafers to be processed are stacked up vertically into slots in a quartz carrier or "boat" and inserted into the furnace lube.

Diffusion Of p-Type Impurity

Boron is an almost exclusive choice as an acceptor impurity in silicon. It has a moderate diffusion coefficient, typically of order I0-16 m2/sec at 1150°C which is convenient for precisely controlled diffusion. It has a solid solubility limit of around 5 x 1026 atoms/m3, so that surface concentration can be widely varied, but most reproducible results are obtained when the concentration is approximately 1024/m3, which is typical for transistor base diffusions.

- **Boron Diffusion using B_2H_6 (Diborane) Source**

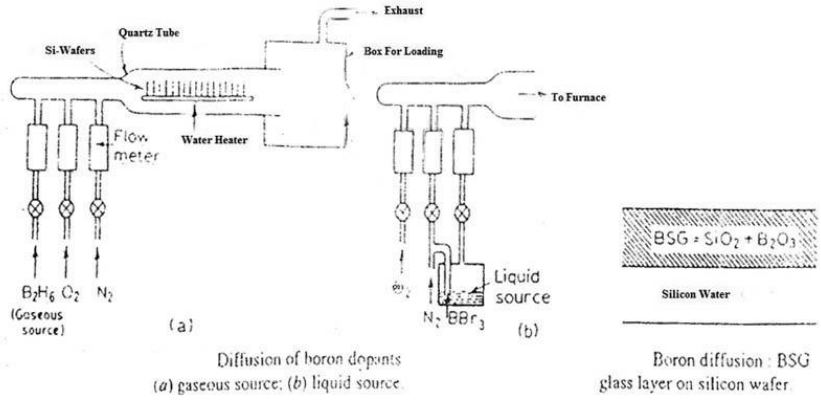
This is a gaseous source for boron. This can be directly introduced into the diffusion furnace. A number of other gases are metered into the furnace. The principal gas flow in the furnace will be nitrogen (N_2) which acts as a relatively inert gas and is used as a carrier gas to be a dilutent for the other more reactive gases. The N_2, carrier gas will generally make up

some 90 to 99 percent of the total gas flow. A small amount of oxygen and very small amount of a source of boron will make up the rest of the gas flow. This is shown in the figure below. The following reactions will be occurring simultaneously at the surface of the silicon wafers:

$Si + 0_2 = SiO_2$ (silica glass)

$2B_2H_6 + 30_2 = B_2O_3$ (boron glass) + $6H_2$

This process is the chemical vapour deposition (CVD) of a glassy layer on (lie silicon surface which is a mixture of silica glass ($Si0_2$) and boron glass (B_20_3) is called borosilica glass (BSG). The BSG glassy layer, shown in the figure below, is a viscous liquid at the diffusion temperatures and the boron atoms can move around relatively easily.



Diffusion of boron dopants (*a*) gaseous source; (*b*) liquid source. Boron diffusion : BSG glass layer on silicon wafer.

Diffusion Of Dopants

Furthermore, the boron concentration in the BSG is such that the silicon surface will be saturated with boron at the solid solubility limit throughout the time of the diffusion process as long as BSG remains present. This is constant source (erfc) diffusion. It is often called deposition diffusion. This diffusion step is referred as pre-deposition step in which the dopant atoms deposit into the surface regions (say 0.3 micro meters depth) of the silicon wafers. The BSG is preferable because it protects the silicon atoms from pitting or evaporating and acts as a "getter" for undesirable impurities in the silicon. It is etched off before next diffusion as discussed below.

The pre-deposition step, is followed by a second diffusion process in which the external dopant source (BSG) is removed such that no additional dopants cuter the silicon. During this diffusion process the dopants that are already in the silicon move further in and are thus redistributed. The junction depth increases, and at the same time the surface

concentration decreases. This type of diffusion is called drive-in, or redistribution, or limited-source (Gaussian diffusion).

- **Boron Diffusion using BBr_3i (Boron Tribromide) Source**

This is a liquid source of boron. In this case a controlled flow of carrier gas (N_2,) is bubbled through boron tribromide, as shown in the figure below, which with oxygen again produces boron trioxide (BSG) at the surface of the wafers as per following reaction :

$4BBr_3 + 30_2 = B20_3 + 2Br_2$

Diffusion of n-Type Impurity

For phosphorus diffusion such compounds as PH3 (phosphine) and POCl3 (phosphorus oxychloride) can be used. In the case of a diffusion using PoCI3, the reactions occurring at the silicon wafer surfaces will be:

Si + 02 = SiO2 (silica glass)

$4POCl + 30_2 = 2P_20_5 + 6Cl_2$

This will result in the production of a glassy layer on the silicon wafers (hat is a mixture of phosphorus glass and silica glass called phosphorosilica glass (PSG), which is a viscous liquid at the diffusion temperatures. The mobility of the phosphorus atoms in this glassy layer and the phosphorus concentration is such that the phosphorus concentration at the silicon surface will be maintained at the solid solubility limit throughout the time of the diffusion process (similar processes occur with other dopants, such as the case of arsenic, in winch arsenosilica glass is formed on the silicon surface.

The rest of the process for phosphorus diffusion is similar to boron diffusion, that is, after deposition step, drive-in diffusion is carried out.P_20_5 is a solid source for phosphorus impurity and can be used in place of $POCl_3$. However $POCl_3$ offers certain advantages overP205 such as easier source handling, simple furnace requirements, similar glassware for low and high surface

CVD Process

A multitude of layers of different materials have to be deposited during the IC fabrication process. The two most important deposition methods are the physical vapor

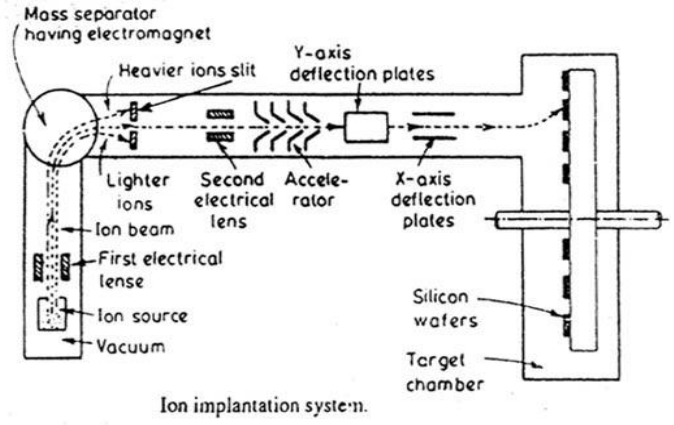
deposition (PVD) and the chemical vapor deposition (CVD). During PVD accelerated gas ions sputter particles from a sputter target in a low pressure plasma chamber. The principle of CVD is a chemical reaction of a gas mixture on the substrate surface at high temperatures. The need of high temperatures is the most restricting factor for applying CVD. This problem can be avoided with plasma enhanced chemical vapor deposition (PECVD), where the chemical reaction is enhanced with radio frequencies instead of high temperatures. An important aspect for this technique is the uniformity of the deposited material, especially the layer thickness. CVD has a better uniformity than PVD.

ION-IMPLANTATION

Ion Implantation is an alternative to a deposition diffusion and is used to produce a shallow surface region of dopant atoms deposited into a silicon wafer. This technology has made significant roads into diffusion technology in several areas. In this process a beam of impurity ions is accelerated to kinetic energies in the range of several tens of kV and is directed to the surface of the silicon. As the impurity atoms enter the crystal, they give up their energy to the lattice in collisions and finally come to rest at some average penetration depth, called the projected range expressed in micro meters. Depending on the impurity and its implantation energy, the range in a given semiconductor may vary from a few hundred angstroms to about 1micro meter. Typical distribution of impurity along the projected range is approximately Gaussian. By performing several implantations at different energies, it is possible to synthesize a desired impurity distribution, for example a uniformly doped region.

Ion Implantation

A typical ion-implantation system is shown in the figure below.



Ion Implantation System

A gas containing the desired impurity is ionized within the ion source. The ions are generated and repelled from their source in a diverging beam that is focussed before if passes through a mass separator that directs only the ions of the desired species through a narrow aperture. A second lens focuses this resolved beam which then passes through an accelerator that brings the ions to their required energy before they strike the target and become implanted in the exposed areas of the silicon wafers. The accelerating voltages may be from 20 kV to as much as 250 kV. In some ion implanters, the mass separation occurs after the ions are accelerated to high energy. Because the ion beam is small, means are provided for scanning it uniformly across the wafers.

Metallization

Metallization is the final step in the wafer processing sequence. Metallization is the process by which the components of IC's are interconnected by aluminium conductor. This process produces a thin-film metal layer that will serve as the required conductor pattern for the interconnection of the various components on the chip. Another use of metallization is to produce metalized areas called bonding pads around the periphery of the chip to produce metalized areas for the bonding of wire leads from the package to the chip. The bonding wires are typically 25 micro meters diameter gold wires, and the bonding pads are usually made to be around 100×100 micro meters square to accommodate fully the flattened ends of the bonding wires and to allow for some registration errors in the placement of the wires on the pads.

Aluminium

Aluminium (At) is the most commonly used material for the metallization of most IC's, discrete diodes, and transistors. The film thickness is as about 1 micro meters and conductor widths of about 2 to 25 micro meters are commonly used. The use of aluminium offers the following advantages:

- It has as relatively good conductivity.
- It is easy to deposit thin films of Al by vacuum evaporation.
- It has good adherence to the silicon dioxide surface.
- Aluminium forms good mechanical bonds with silicon by sintering at about 500°C or by alloying at the eutectic temperature of 577°C.

- Aluminium forms low-resistance, non-rectifying (that is, ohmic) contacts with p-type silicon and with heavily doped n-type silicon.
- It can be applied and patterned with a single deposition and etching process.

In general the desired properties of the metallization for IC can be listed as follows.

- Low resistivity.
- Easy to form.
- Easy to etch for pattern generation.
- Should be stable in oxidizing ambient , oxidizable.
- Mechanical stability; good adherence, low stress.
- Surface smoothness.
- Stability throughout processing including high temperature sinter, dry or wet oxidation, gettering, phosphorous glass (or any other material) passivation, metallization.
- No reaction with final metal, aluminium.
- Should not contaminate device, wafers, or working apparatus.
- Good device characteristics and life times.
- For window contacts-low contact resistance, minimum junction penetration, low electromigration.

Threshold Voltage and Body Effect

The threshold voltage V_{th} for a nMOS transistor is the minimum amount of the gate-to-source voltage V_{GS} necessary to cause surface inversion so as to create the conducting channel between the source and the drain. For $V_{GS} < V_{th}$, no current can flow between the source and the drain. For $V_{GS} > V_{th}$, a larger number of minority carriers (electrons in case of an nMOS transistor) are drawn to the surface, increasing the channel current. However, the surface potential and the depletion region width remain almost unchanged as V_{GS} is increased beyond the threshold voltage.

The physical components determining the threshold voltage are the following.

- work function difference between the gate and the substrate.
- gate voltage portion spent to change the surface potential.
- gate voltage part accounting for the depletion region charge.

- gate voltage component to offset the fixed charges in the gate oxide and the silicon-oxide boundary.

Although the following analysis pertains to an nMOS device, it can be simply modified to reason for a p-channel device.

The work function difference ϕ_{GS} between the doped polysilicon gate and the p-type substrate, which depends on the substrate doping, makes up the first component of the threshold voltage. The externally applied gate voltage must also account for the strong inversion at the surface, expressed in the form of surface potential $2\phi_F$, where ϕ_F denotes the distance between the intrinsic energy level E_I and the Fermi level E_F of the p-type semiconductor substrate.

The factor 2 comes due to the fact that in the bulk, the semiconductor is p-type, where E_I is above E_F by ϕ_F, while at the inverted n-type region at the surface E_I is below E_F by ϕ_F, and thus the amount of the band bending is $2\phi_F$. This is the second component of the threshold voltage. The potential difference ϕ_F between E_I and E_F is given as

$$\phi_F = \frac{kT}{q} \ln\left(\frac{N_A}{n_i}\right)$$

where k: Boltzmann constant, T: temperature, q : electron charge N_A : acceptor concentration in the p-substrate and n i : intrinsic carrier concentration. The expression kT/q is 0.02586 volt at 300 K.

The applied gate voltage must also be large enough to create the depletion charge. Note that the charge per unit area in the depletion region at strong inversion is given by

$$Q_{d0} = -2(\varepsilon_S q N_A \phi_F)^{1/2}$$

where ε_S is the substrate permittivity. If the source is biased at a potential V_{SB} with respect to the substrate, then the depletion charge density is given by

$$Q_d = -2(\varepsilon_S q N_A (\phi_F + V_{SB}))^{1/2}$$

The component of the threshold voltage that offsets the depletion charge is then given by $-Q_d$ $/C_{ox}$, where C_{ox} is the gate oxide capacitance per unit area, or $C_{ox} = \varepsilon_{ox} / t_{ox}$ (ratio of the oxide permittivity and the oxide thickness).

A set of positive charges arises from the interface states at the $Si\text{-}SiO_2$ interface. These charges, denoted as Q_i , occur from the abrupt termination of the semiconductor crystal lattice at the oxide interface. The component of the gate voltage needed to offset this positive charge (which induces an equivalent negative charge in the semiconductor) is $-Q_i / C_{ox}$. On combining all the four voltage components, the threshold voltage V_{TO}, for zero substrate bias, is expressed as

$$V_{T0} = \phi_{GS} - 2\phi_F - \frac{Q_{d0}}{C_{ox}} - \frac{Q_i}{C_{ox}}$$

For non-zero substrate bias, however, the depletion charge density needs to be modified to include the effect of V_{SB} on that charge, resulting in the following generalized expression for the threshold voltage, namely

$$V_T = \phi_{GS} - 2\phi_F - \frac{Q_d}{C_{ox}} - \frac{Q_i}{C_{ox}}$$

The generalized form of the threshold voltage can also be written as

$$V_T = \phi_{GS} - 2\phi_F - \frac{Q_{d0}}{C_{ox}} - \frac{Q_i}{C_{ox}} - \frac{Q_d - Q_{d0}}{C_{ox}} = V_{T0} - \frac{Q_d - Q_{d0}}{C_{ox}}$$

Note that the threshold voltage differs from V_{TO} by an additive term due to substrate bias. This term, which depends on the material parameters and the source-to-substrate voltage V_{SB} , is given by

$$\frac{Q_d - Q_{d0}}{C_{ox}} = -\frac{\sqrt{2qN_A\varepsilon_S}}{C_{ox}}\left(\sqrt{|2\phi_F + V_{SB}|} - \sqrt{|2\phi_F|}\right)$$

Thus, in its most general form, the threshold voltage is determined as

$$V_T = V_{T0} + \gamma\left(\sqrt{|2\phi_F + V_{SB}|} - \sqrt{|2\phi_F|}\right) \quad \text{..... (2.1)}$$

in which the parameter γ, known as the *substrate-bias* **(or** *body-effect* **)** coefficient is given by

$$\gamma = \frac{\sqrt{2qN_A\varepsilon_S}}{C_{ox}} \quad \text{..... (2.2)}$$

The threshold voltage expression given by (1.1) can be applied to n-channel as well as p-channel transistors. However, some of the parameters have opposite polarities for the pMOS and the nMOS transistors. For example, the substrate bias voltage V_{SB} is positive in nMOS and negative in pMOS devices. Also, the substrate potential difference ϕ_F is negative in nMOS, and positive in pMOS. Whereas, the body-effect coefficient γ is positive in nMOS and negative in pMOS. Typically, the threshold voltage of an enhancement mode n-channel transistor is positive, while that of a p-channel transistor is negative.

Example 2.1 Given the following parameters, namely the acceptor concentration of p-substrate $N_A = 10^{16}$ cm^{-3} , polysilicon gate doping concentration $N_D = 10^{16}$ cm^{-3} , intrinsic concentration of Si, $n_i = 1.45 \times 10^{10}$ cm^{-3} , gate oxide thickness $t_{ox} = 500$ Å and oxide-interface fixed charge density $N_{ox} = 4 \times 10^{10}$cm^{-2} , calculate the threshold voltage V_{TO} at $V_{SB}=0$.

Ans:

The potential difference between E_I and E_F for the p-substrate is

$$\phi_F = KT/q \ln\left(N_A/n_i\right) = 0.026V\, C_n\left(10^{16}/1.45\times10^{10}\right) = 0.35V$$

For the polysilicon gate, as the doping concentration is extremely high, the heavily doped n-type gate material can be assumed to be degenerate. That is, the Fermi level E_F is almost coincident with the bottom of the conduction band E C . Hence, assuming that the intrinsic energy level E_I is at the middle of the band gap, the potential difference between E_I and E_F for the gate is ϕ_F = ½ (energy band gap of Si) = 1/2 X 1.1 = 0.55 V.

Thus, the work function difference ϕ_{GS} between the doped polysilicon gate and the p-type substrate is -0.35 V - 0.55 V = -0.90 V.

The depletion charge density at $V_{SB} = 0$ is

$$Q_{d0} = -2\left(\varepsilon_S q N_A \phi_F\right)^{1/2} = -2\left(11.7\times8.85\times10^{-14}\times1.6\times10^{-19}\times10^{16}\times0.35\right)^{1/2} = -4.82\times10^{-8}\, C/cm^2$$

The oxide-interface charge density is

$$Q_i = q.N_{ox} = 1.6\times10^{-19}\ \text{C} \times 4 \times10^{10}\text{cm}^{-2} = 6.4\times10^{-19}\ \text{C/m}^2$$

The gate oxide capacitance per unit area is (using dielectric constant of SiO_2 as 3.97)

$$C_{ox} = \varepsilon_{ox}/t_{ox} = \left(3.97\times8.85\times10^{-14}\ \text{F/cm}\right)/\left(500\times10^{-8}\ \text{cm}\right) = 7.03\times10^{-8}\ \text{F/cm}^2$$

Combining the four components, the threshold voltage can now be computed as

$$V_{TO} = \phi_{GS} - 2\phi_F(\text{substrate}) - Q_{do}/C_{OX} - Q_i/C_{OX} = -0.90 - (-0.69) - 0.09 = 0.40 \text{ volt}$$

Body Effect :

The transistors in a MOS device seen so far are built on a common substrate. Thus, the substrate voltage of all such transistors are equal. However, while one designs a complex gate using MOS transistors, several devices may have to be connected in series. This will result in different source-to-substrate voltages for different devices. For example, in the NAND gate shown in Figure 1.5, the nMOS transistors are in series, whereby the source-to-substrate voltage V_{SB} of the device corresponding to the input A is higher than that of the device for the input B.

Under normal conditions ($V_{GS} > V_{th}$), the depletion layer width remains unchanged and the charge carriers are drawn into the channel from the source. As the substrate bias V_{SB} is increased, the depletion layer width corresponding to the source-substrate field-induced junction also increases. This results in an increase in the density of the fixed charges in the depletion layer. For charge neutrality to be valid, the channel charge must go down. The consequence is that the substrate bias V_{SB} gets added to the channel-substrate junction potential. This leads to an increase of the gate-channel voltage drop.

Example 2.2 Consider the n-channel MOS process in Example 2.1. One may examine how a non-zero source-to-substrate voltage V_{SB} influences the threshold voltage of an nMOS transistor.

One can calculate the substrate-bias coefficient γ using the parameters provided in Example 2.1 as follows :

$$\gamma = \frac{\sqrt{2qN_A\varepsilon_S}}{C_{ox}} = \frac{\sqrt{2*1.6*10^{-19}*10^6*11.7*8.85*10^{-14}}}{7.03*10^{-18}} = 0.82\ V^{\frac{1}{2}}$$

One is now in a position to determine the variation of threshold voltage V_T as a function of the source-to-substrate voltage V_{SB} . Assume the voltage V_{SB} to range from 0 to 5 V.

$$V_T = V_{T0} + \gamma\left(\sqrt{|2\phi_F + V_{SB}|} - \sqrt{|2\phi_F|}\right) = 0.40 + 0.82(\sqrt{0.7 + V_{SB}} - \sqrt{0.7})$$

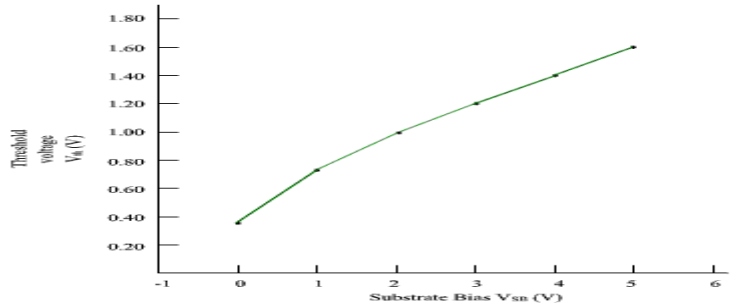


Figure 2.7 Variation of Threshold voltage in response to change in source-to-substrate voltage V_{SB}

Figure 2.7 depicts the manner in which the threshold voltage V_{th} varies as a function of the source-to-substrate voltage V_{SB} . As may be seen from the figure, the extent of the variation of the threshold voltage is nearly 1.3 Volts in this range. In most of the digital circuits, the substrate bias effect (also referred to as the body effect) is inevitable. Accordingly, appropriate measures have to be adopted to compensate for such variations in the threshold voltage.

INTRODUCTION:

The basic characteristics of MOS transistor and the various possibilities of configuring inverter circuits are explained in this unit BiCMos transistors are also considered.

The expressions and discussion is done for nMos transistors and the PMOS expressions given polarities of nMos expressions. The exchange of μ_n for μ_p and electrons for holes will result in PMOS from η MOS expressions.

DRAIN TO SOURCE CURRENT I_{ds} VERSUS VOLTAGE V_{ds} RELATIONSHIPS:-

When a voltage is applied on the gate of a MOS transistor a charge is induced in the channel between source and drain. This charge moves from source to drain when a voltage V_{ds} is applied between drain and source. I_{ds} is dependent on V_{gs} and V_{ds}

$$I_{ds} = -I_{sd} = \frac{charge induced in channel (QC)}{Electron transit time (T)} \quad \text{_____(1)}$$

$$\text{Transit time } \tau_{sd} = \frac{Length of channel (L)}{velocity (v)} \quad \text{_____(2)}$$

Where $v = \mu E_{ds}$

μ = Electron or hole mobility

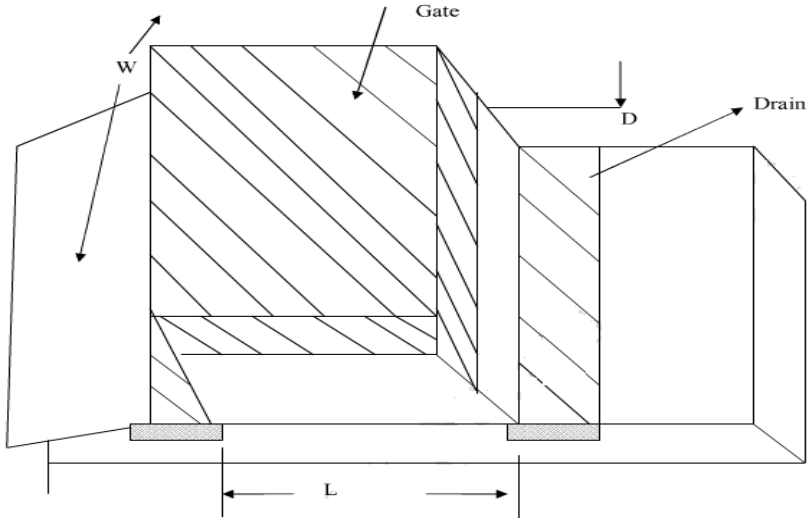
E_{ds} = electric field (drain to source)

We have $E_{ds} = \frac{V_{ds}}{L}$

So $v = \frac{\mu V_{ds}}{L}$

Substituting v in (2)

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



Derivation for I_{ds}:-

1. Non-Saturated Region:-

As in non saturated region

$V_{ds} < V_{gs} - V_t$

As said in the previous section charge is induced in the channel and we have

Charge/unit area=Eg $\varepsilon ins \varepsilon_o$

$$Q_c = Eg\, \varepsilon ins \varepsilon_o . \frac{WL}{unitarea} \quad \text{_____(3)}$$

εins =relative permittivity of insulation between gate and channel

ε_o = permittivity of free space

ε_o = 8.85 x 10^{-14}Fcm^{-1}

We have Eg = $\dfrac{\left[\left(V_{gs}-V_t\right)-\dfrac{V_{ds}}{2}\right]}{D}$

D = onide thickness

$\therefore$ Substituting Eg in (3) we get

$$Q_c = \frac{WL\varepsilon ins\varepsilon_o}{D}\left[\left(V_{gs}-V_t\right)-\frac{V_{ds}}{2}\right]$$

Substituting Q_c in equation (1)-$I_{ds} = \dfrac{Q_c}{\tau}$

We get

$$I_{ds} = \frac{\varepsilon ins\varepsilon_o\mu}{D}\frac{W}{L}\left[\left(V_{gs}-V_t\right)-\frac{V_{ds}}{2}\right]V_{ds}$$

$$= k.\frac{W}{L}\left[\left(V_{gs}-V_t\right)V_{ds}-\frac{V_{ds}^{\,2}}{2}\right]$$

We take $\beta = K.\dfrac{W}{L}$

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

We also have gate channel capacitance

$$Cg = \frac{\varepsilon ins\varepsilon_o.W.L}{D}$$

So $K = \dfrac{Cg.\mu}{WL}$

$$\therefore I_{ds} = \frac{Cg\mu}{L^2}\left[\left(V_{gs}-V_t\right)V_{ds}-\frac{V_{ds}^{\,2}}{2}\right]$$

2. In saturated region we have

$V_{ds} = V_{gs}$-V

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

$$= \frac{Cg\mu}{2L^2}(V_{gs} - V_t)^2$$

MOS TRANSISTOR THRESHOLD VOLTAGE V_t:-

The charges stored in the dielectric layers and in the surface to surface interfaces ar neutralized by switching and enhancement mode transistor from off to on by applyi sufficient gate voltage

The threshold voltage V_t may be expressed as

$$V_t = \phi_{ms} \frac{Q_B - Q_{ss}}{C_o} + 2\phi_{fN}$$

Q_B = charge per unit area in the depletion layer beneath the oxide

Q_{ss} = charge density at Si+SiO_2 interface.

C_o = Capacitance per unit gate area

Q_{ms} = Work function difference gate and Si

Q_{fN} = Fermi level potential between inverted surface and bulk Si.

For PolySilicon gate and Silicon substrate the is determined as follows

$$Q_B = \sqrt{2\varepsilon_o \varepsilon_s qN(2\phi_{fN} + V_{SB})} coloumb/m^2$$

$$\phi_{fN} = \frac{KT}{q} \ln \frac{N}{ni} volts$$

$Q_{ss} = (1.5x10^8)x10^{-8}$colomb/$m^2$

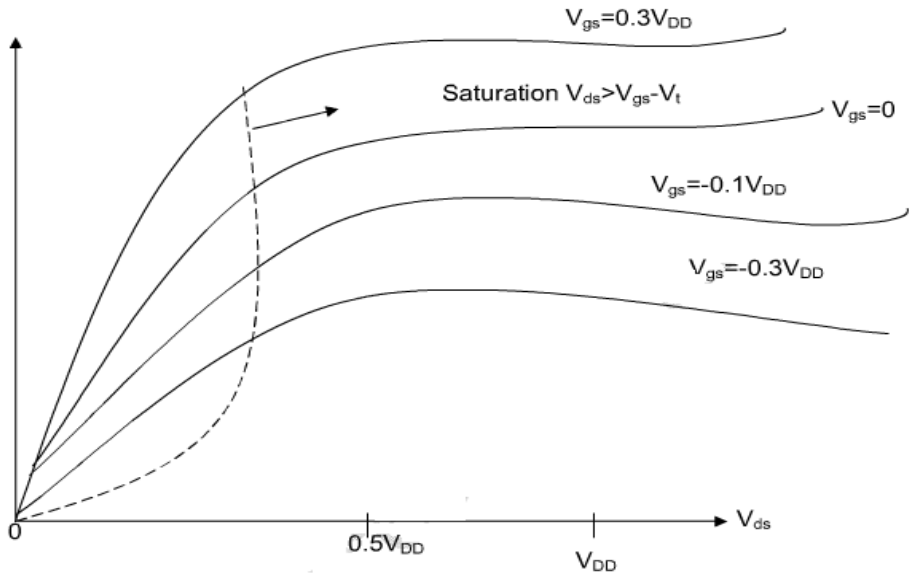
Where V_{SB} = Substrate bias voltage, q=$1.6x10^{-19}$coloumb,

N – impurity permittivity of silicon

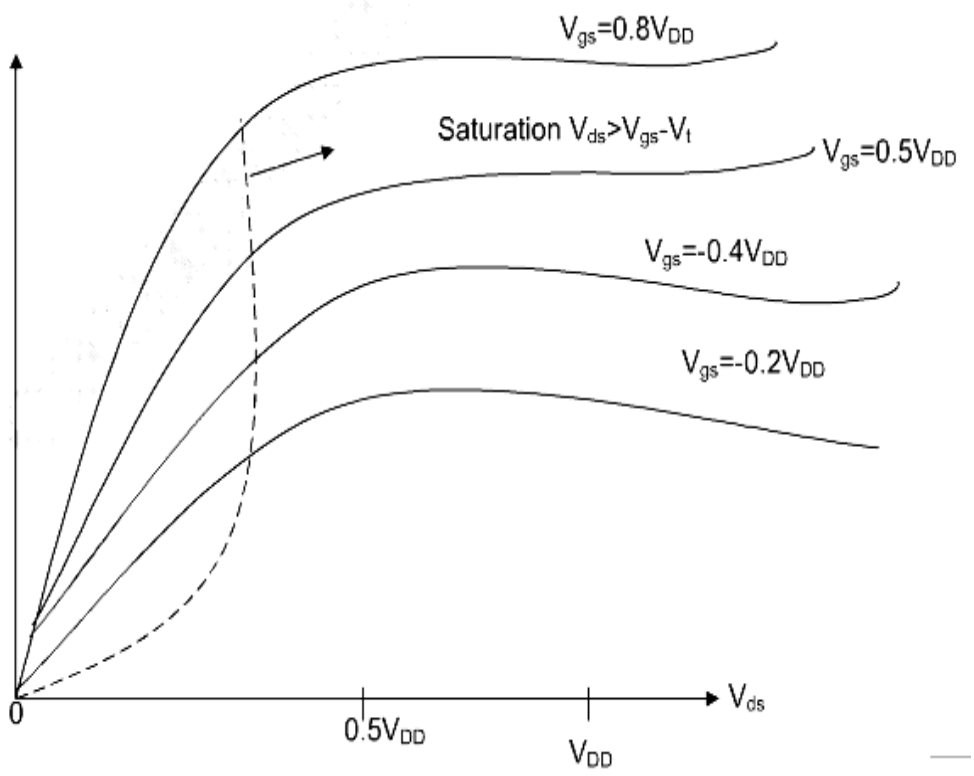
ε_{Si} – relative permittivity of Silicon

MOS transistor Characteristics

(a) Depletion mode device:-

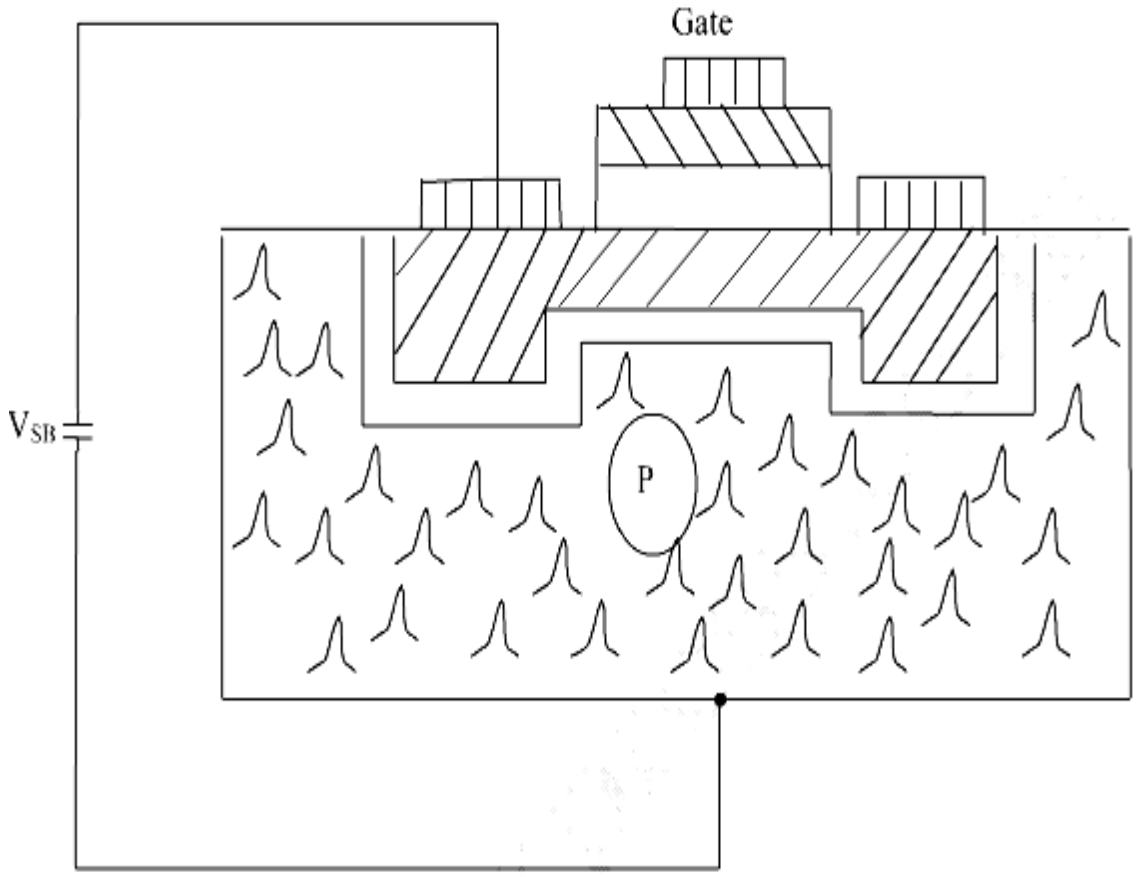


(b)Enhancement mode device:



BODY EFFECT

The substrate may be biased with respect to the source



Body effect (nMOS device shown)

When V_{SB} is increased, the channel is depleted of charge carriers and thus the threshold voltage is raised
Threshold is give by voltage

$V_t = \gamma(V_{SB})^{\frac{1}{2}}$ where γ is a constant which depends on substrate doping. The body effect will be smaller if the substrate is doped lightly

It also be written as

$$V_t = V_t(0) + \left[\frac{D}{\varepsilon ins\,\varepsilon_o}\right]\sqrt{2\varepsilon_o \varepsilon SiQN}\,(V_{SB})^{\frac{1}{2}}$$

Where $V_t(0)$ is the threshold voltage for $V_{SB}=0$

MOS TRANSISTOR TRANSCONDUCTANCE gm

Trans conductance expresses the relationship between output current I_{ds} and input voltage V_{gs} and is defined as

$$I_m = \frac{\delta I_{ds}}{\delta V_{gs}} \Big|_{V_{ds}=constant}$$

We have $\frac{Q_c}{I_{ds}} = \tau$

$$\therefore \delta I_{ds} = \frac{\delta Q_c}{\tau_{ds}}$$

Substituting $\tau_{ds} = \frac{L^2}{\mu V_{ds}}$ in above equation

$$\delta I_{ds} = \frac{\delta Q_c . \mu . V_{ds}}{L^2}$$

But $\delta Q_c = Cg . \delta V_{gs}$

So $\delta I_{ds} = \frac{C_g \delta V_{gs} . \mu . V_{ds}}{L^2}$

Now $g_m = \frac{\delta I_{ds}}{\delta V_{gs}} = \frac{Cg . \mu . V_{ds}}{L^2}$

In saturation $V_{ds} = V_{gs} - V_t$

$$g_m = \frac{C_g . \mu}{L^2}(V_{gs} - V_t)$$

Substituting $C_g = \frac{\varepsilon ins \varepsilon_o WL}{D}$

$$gm = \frac{\mu \varepsilon \sin \varepsilon_o}{D} . \frac{W}{L}(V_{gs} - V_t)$$

$$\Rightarrow gm = \beta(V_{gs} - V_t)$$

Output CONDUCTANCE g_{ds}

It can be expressed as

$$g_{ds} = \frac{\delta I_{ds}}{\delta V_{gs}} = \lambda . I_{ds} \propto \left(\frac{1}{L}\right)^{Z}$$

$$\lambda \propto \frac{1}{L} and I_{ds} \propto \frac{1}{L} for a \text{ MOS device.}$$

MOS TRANSISTOR FIGURE OF MERIT

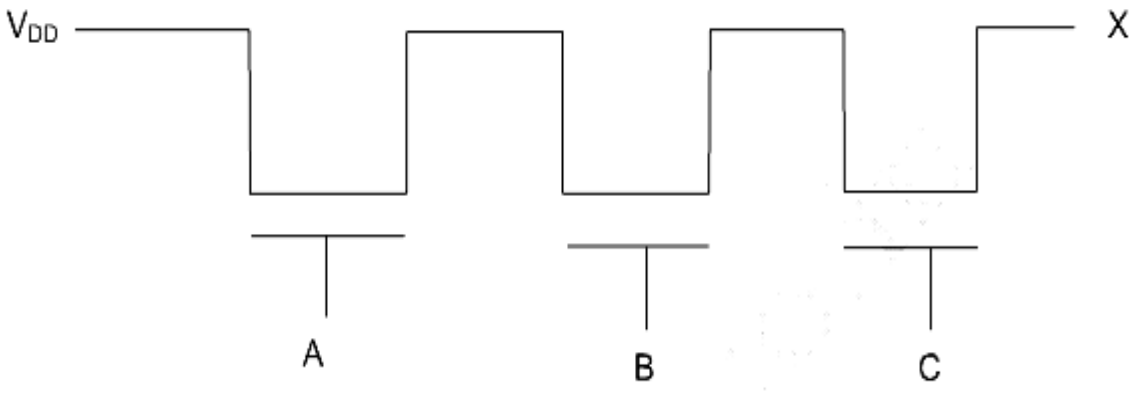
The parameter w_o is useful to find the frequency response.

$$w_o = \frac{g_m}{C_g} = \frac{\mu}{L^2}(V_{gs} - V_t)\left[= \frac{1}{\tau_{sd}}\right]$$

gm should be as high as possible for a fast circuit. Switching speed is determined by w_o

PASS TRANSISTOR

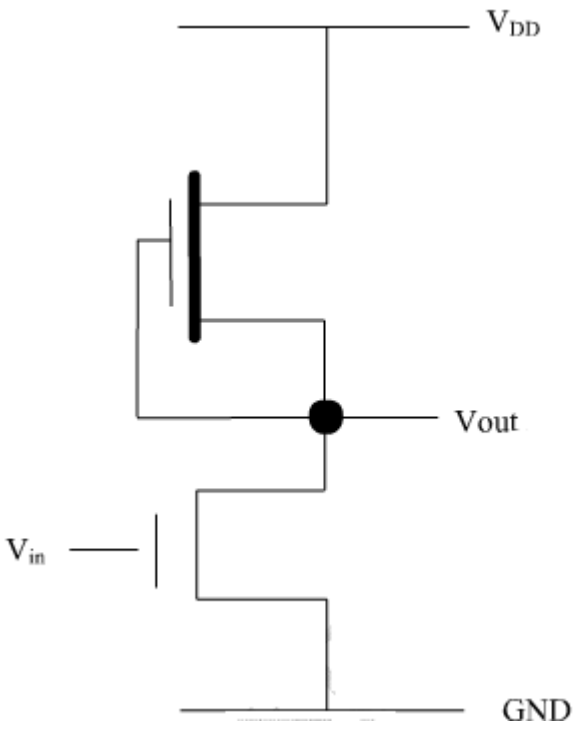
MOS transistors can be used as switches in series. This application is called as pass transistor. The and away is shown below



X = A.B.C (Logic 1= V_{DD}-V_t)

NMOS INVERTER

The inverter circuit consists of transistor with source connected to ground and a depletion mode transistor acting as a load resistor, connected between Drain and V_{DD}.



NMOS inverter

We have

The depletion mode transistor always on because gate is connected to source

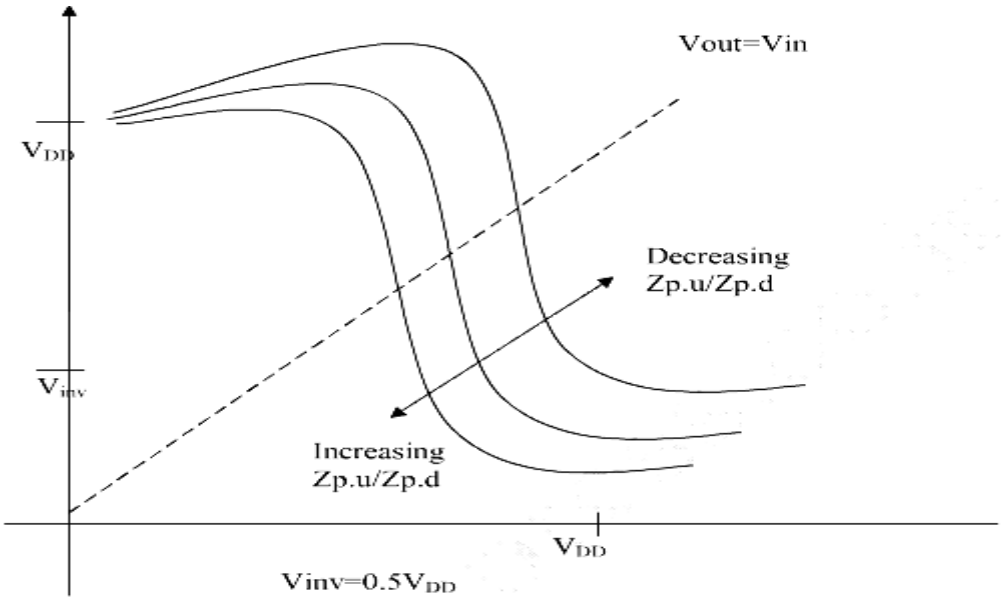
The depletion mode transistor is called pull up device and enhancement mode device is called pull down transistor

The inverter transfer characteristic is obtained by superimposing V_{gs}=0 depletion mode characteristic curve on family of waves for enhancement mode device

The points of intersection give the transfer characteristic

When V in exceeds the p.d threshold voltage current begin to flow

vout



NMOS inverter transfer characterstic

Slope of the transfer characteristic determines the gain

$$\text{Gain} = \frac{\delta Vout}{\delta Vin}$$

PULL UP TO PULL DOWN RATIO (Zp.u/Zp.d) FOR nMOS INVERTER DRIVEN BY ANOTHER nMOS INVERTER

Consider the cascaded inverters in the figure below.

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

At a point $V_{inv} = 0.5V_{DD}$ both the transistors are in saheration and

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

In the depletion mode

$$I_{ds} = K.\frac{W_{p.u}}{L_{p.u}}\frac{(-V_{td})^2}{2} \sin ce V_{gs} = 0$$

And in enhancement mode

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

Equating (since currents are 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 $W_{p.d}$, $L_{p.d}$, $W_{p.a}$ and $L_{p.u}$ are widths and lengths of pull down and pull up transistors respectively.

Now

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

We have

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

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

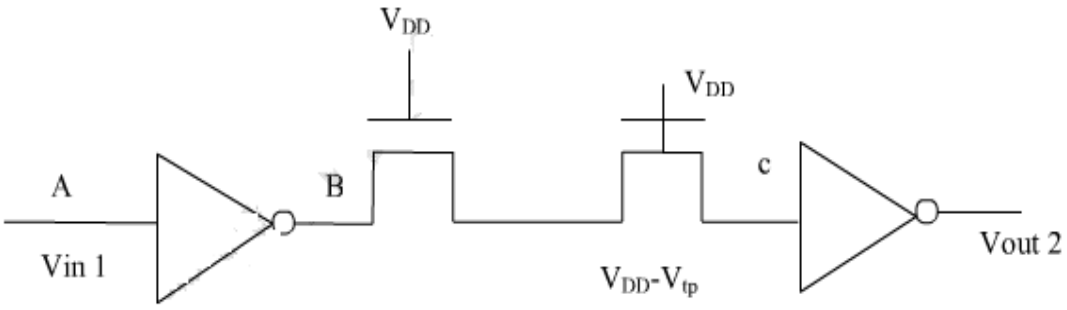
The typical values $V_t = 0.2V_{DD}$, $V_{td}=0.6V_{DD}$ and $V_{inv}=0.5V_{DD}$ are substituted in the above equation

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

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

$$\frac{Z_{p.u}}{Z_{p.d}} = \frac{4}{1}$$

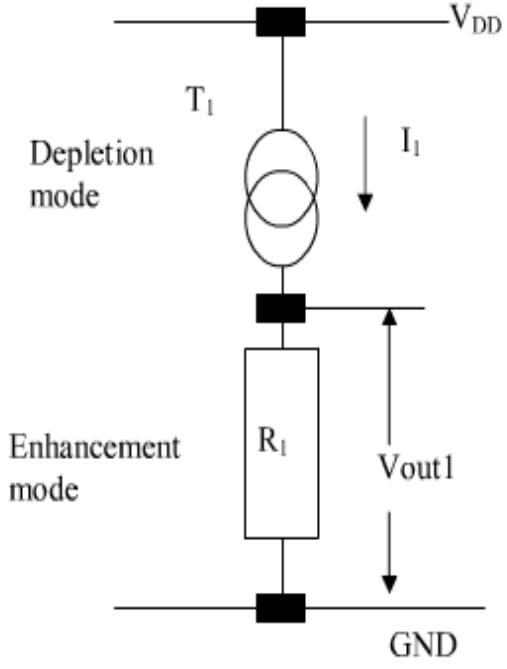
Pull up to Pull Down Ratio for an nMOS Inverter Driven Through One or More Pass Transistors:-



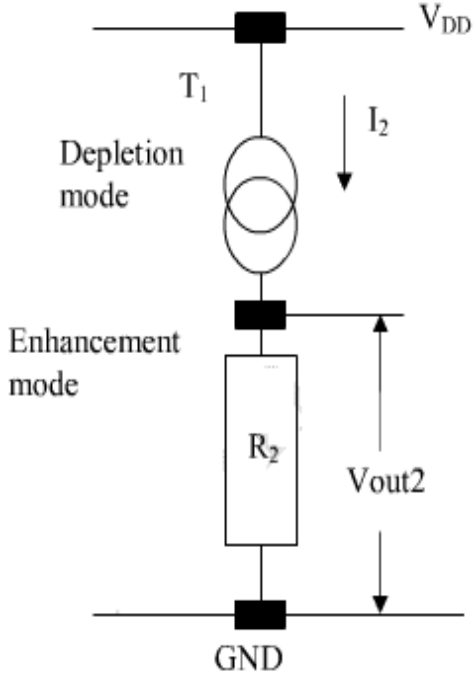
The connection of pass transistors will degrade the logic 1 level into inverter 2 so that the output 0 will not be a proper logic 0 level

When V_{int}=0 voltage at B = V_{DD}

And Input at C = V_{in2} = V_{DD}-V_{tp}
Where V_{tp} = threshold voltage for a pass transistor.



(a) Inverter 1 with input = V_{DD}



(b) Inverter 2 with Input = V_{DD}-V_{tp}

Equivalent Circuits of inverters 1 and 2

For the p.d transistor

$$I_{ds} = K.\frac{W_{p.d1}}{L_{p.d1}}\left[(V_{DD} - V_t)V_{ds1} - \frac{V^2_{ds1}}{2}\right]$$

Therefore

$$R_1 = \frac{V_{ds1}}{I_{ds}} = \frac{1}{K}.\frac{L_{p.d1}}{W_{p.d1}}\left[\frac{1}{V_{DD} - V_t - \frac{V_{ds1}}{2}}\right]$$

For depletion mode transistor in saturation with $V_{gc} = ($

$$I_1 = I_{ds} = K.\ \frac{W_{p.u1}}{L_{p.u1}}\frac{(-V_{td})^2}{2}$$

The product $I_1R_1 = V_{out}1$

$$Vout1 = I_1R_1 = \frac{Zp.d1}{Zp.u.1}\left[\frac{1}{V_{DD} - V_t}\right]\frac{(-V_{td})^2}{2}$$

When input = V_{DD}-V_{tp}

$$R_2 = \frac{1}{K}.Zp.dZ\frac{1}{(V_{DD} - V_{tp}) - V_t}$$

$$I_2 = K.\ \frac{1}{Zp.uZ}\frac{(-V_{td})^2}{2}$$

$$Vout2 = I_2R_2 = \frac{Zp.d2}{Zp.u2}\left[\frac{1}{V_{DD} - V_{tp} - V_t}\right]\frac{(-V_{td})^2}{2}$$

The necessary condition is

Vout1 = Vout2

There fore

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

Taking typical values

$V_t = 0.2V_{DD}$
$V_{tp} = 0.3V_{DD}$

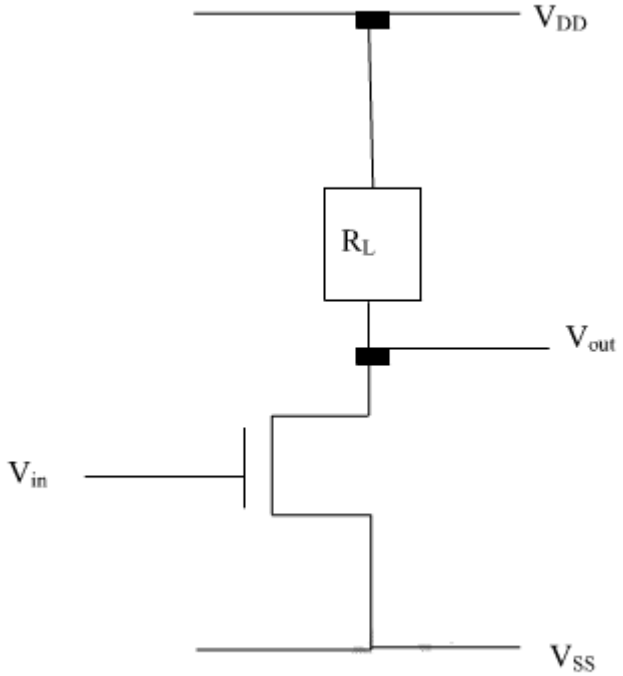
$$\frac{Zp.u.2}{Zp.d.2} = \frac{Zp.u.1}{Zp.d.1}.\frac{0.8}{0.5}$$

$$\frac{Zp.u.2}{Zp.d.2} \approx 2.\frac{Zp.u.1}{Zp.d.1} = 2\frac{4}{1} = \frac{8}{1} \text{ (Approximately equal)}$$

ALTERNATIVE FORMS OF PULL-UP

Load-Resistance RL

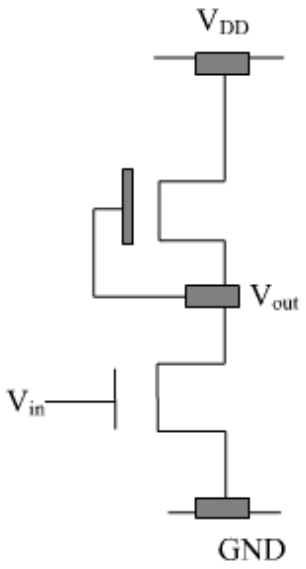
Resistors take large space. So they are not often used.

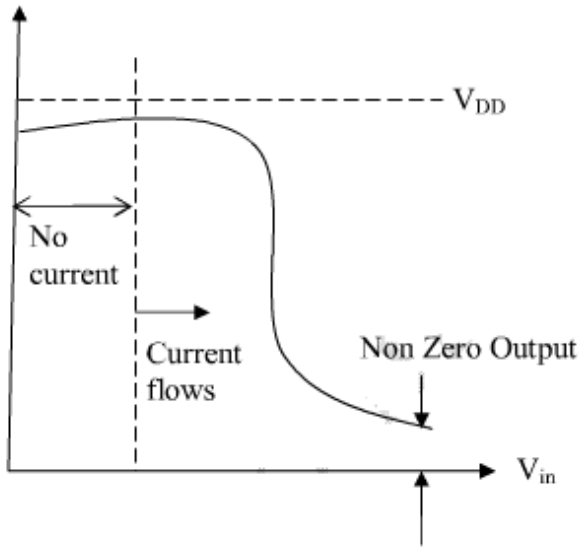


Resistor Pull-up

nMOS Deletion Mode Transistor Pull-Up

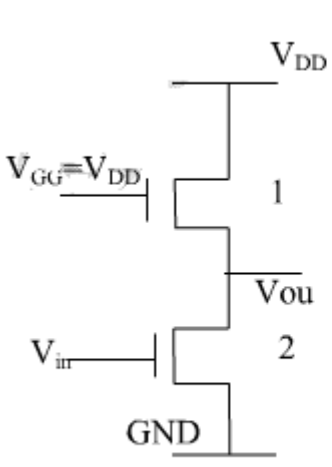
(a) Dissination is high when V_{in} = logical,
(b) Switch of output from 1 to 0 begins when V_{in} exceeds V_t of p.d device

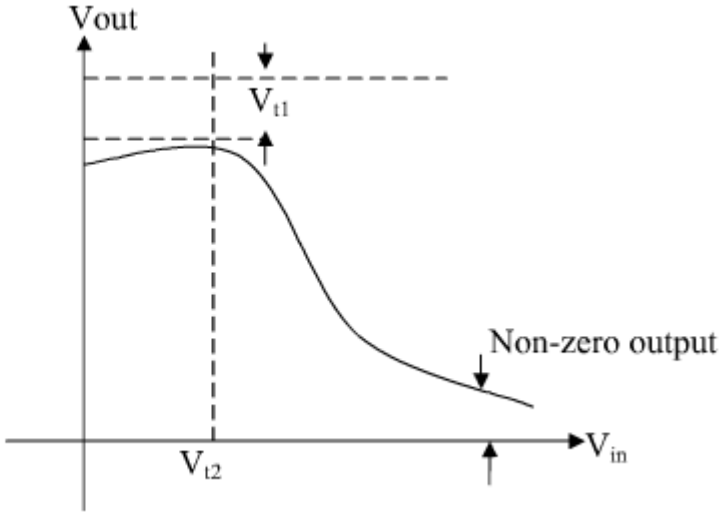




nMOS Enhancement Mode Pull-Up

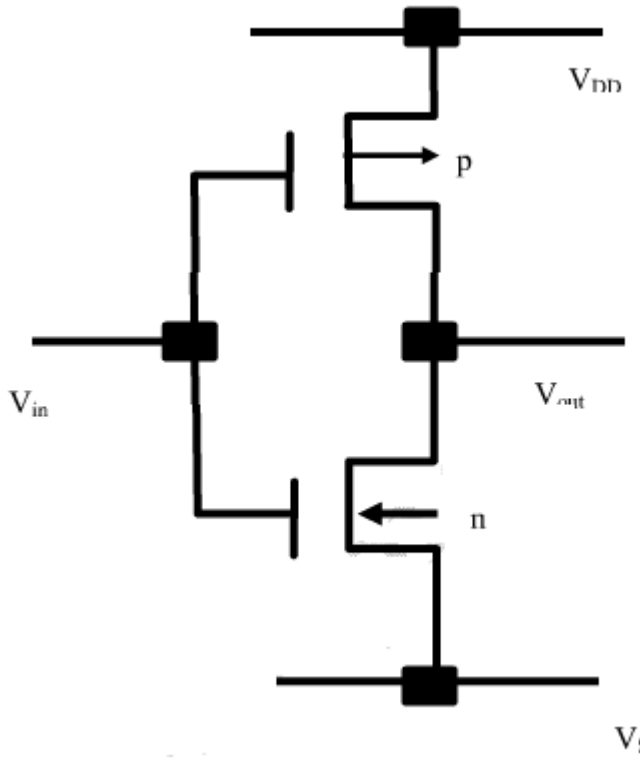
(a) Dissipation is high since current flows when V_{in} = logical 1 (V_{GG} is returned to V_{DD}).
(b) V_{out} can never reach V_{DD}.

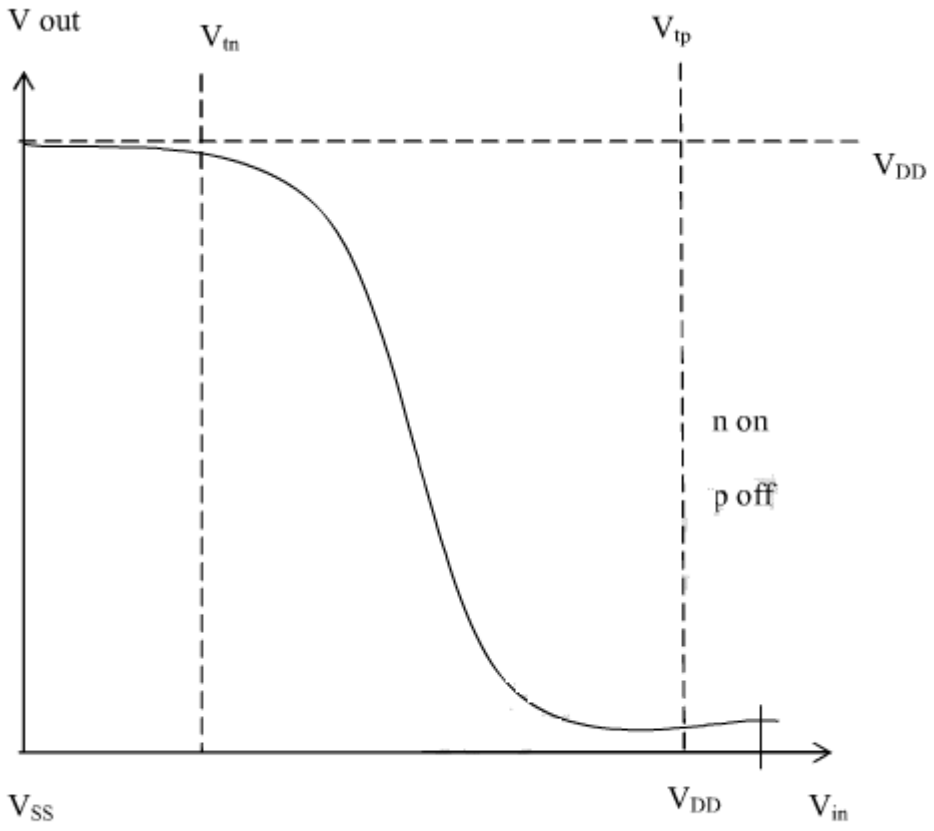




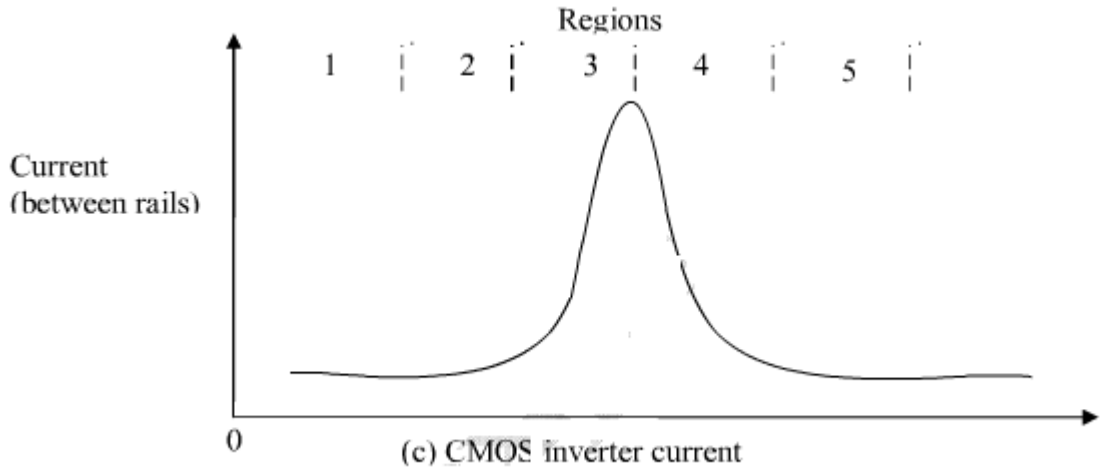
Complementary Transistor Pull-up (CMOS)

(a) No current flow for both logical 0 and 1 inputs.
(b) Iull logical 1 and 0 levels are presented at output.





(b) Transfer Characteristic



(c) CMOS inverter current

Complementary transistor pull-up

CMOS INVERTER ANALYSIS AND DESIGN

The connection and characteristics are shown in above diagram. In the resistive region

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

In the saturation region

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

$$B = K\frac{W}{L}$$

β may be applied to both nMOS and p-MOS transistors

$$\beta_n = \frac{\varepsilon ins\varepsilon_o\mu_n}{D}.\frac{W_n}{L_n}$$

$$\beta_p = \frac{\varepsilon ins\varepsilon_o\mu_p}{D}.\frac{W_p}{L_p}$$

In region 5, V_{in} = 1 and Vout = 0. In region 1 V_{in}=logic 0 and V_{out}=1.

Here p-transistor is fully turned ON and n-transistor is fully turned off.

In region 2, the analysis is done by equating p-device resistive region current with n-device saturation current. Region 4 is similar to region 2 with the functions of p and n transistors reversed.

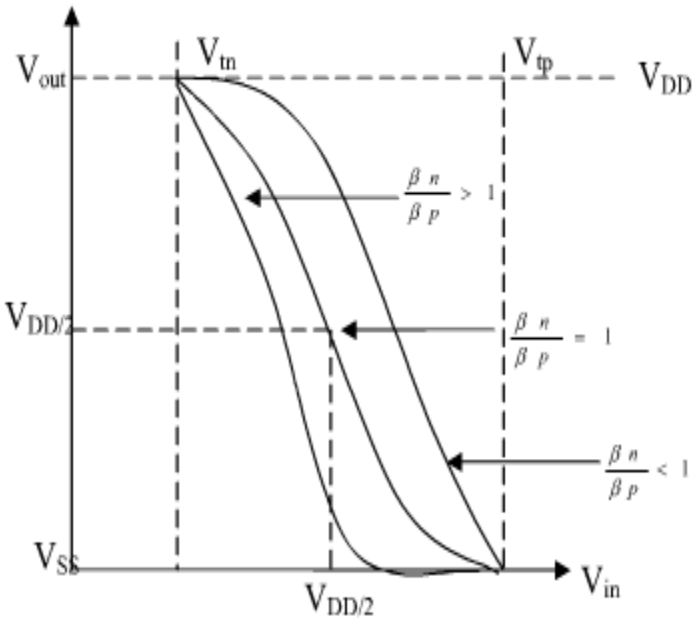
The currents must be the same in each device, so

$$I_{dsp} = -I_{dsn}$$

$$\frac{B_p}{Z}(V_{in} - V_{DD} - V_{tp})^2 = \frac{\beta_n}{2}(V_{in} - V_{tn})^2$$

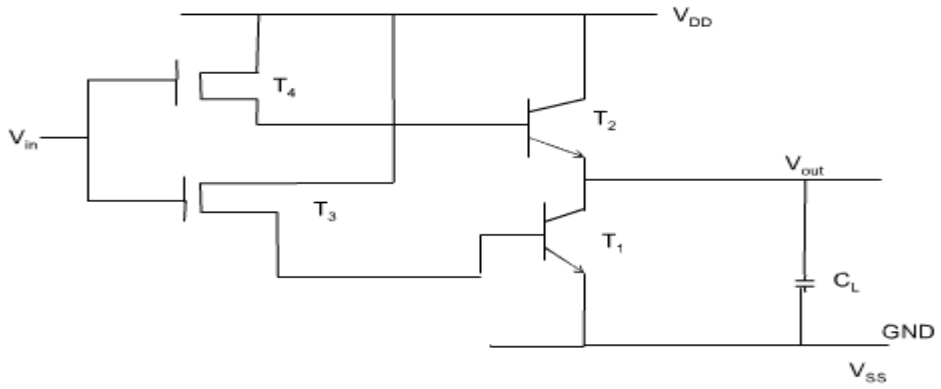
$$\therefore V_{in} = \frac{V_{DD} + V_{tp} + V_{in}\left(\beta_n / \beta_p\right)^{\frac{1}{2}}}{1 + \left(\beta_n / \beta_p\right)^{1/2}}$$

The transfer characteristic with β ratio is shown in the figure below.



BI-CMOS INVERTERS

A simple BiCMOS inverters circuit is shown in the following figure



Bipolar transistors are used to drive the output loads

When $V_{in} = 0$, T_3 and T_1 all off, T_4 and T_2 are on and $V_{out} = +5V$

When $V_{in} = +5V$, T_4 and T_2 are off and T_3 and T_1 are on so the capacitor C_L is discharged to o volts

Characteristics

The inverter has high input impedance
The inverter has low output impedance
It has high noise margins
It occupies relatively small area.

UNIT-II

VLSI CIRCUIT DESIGN PROCESSES

Contents:

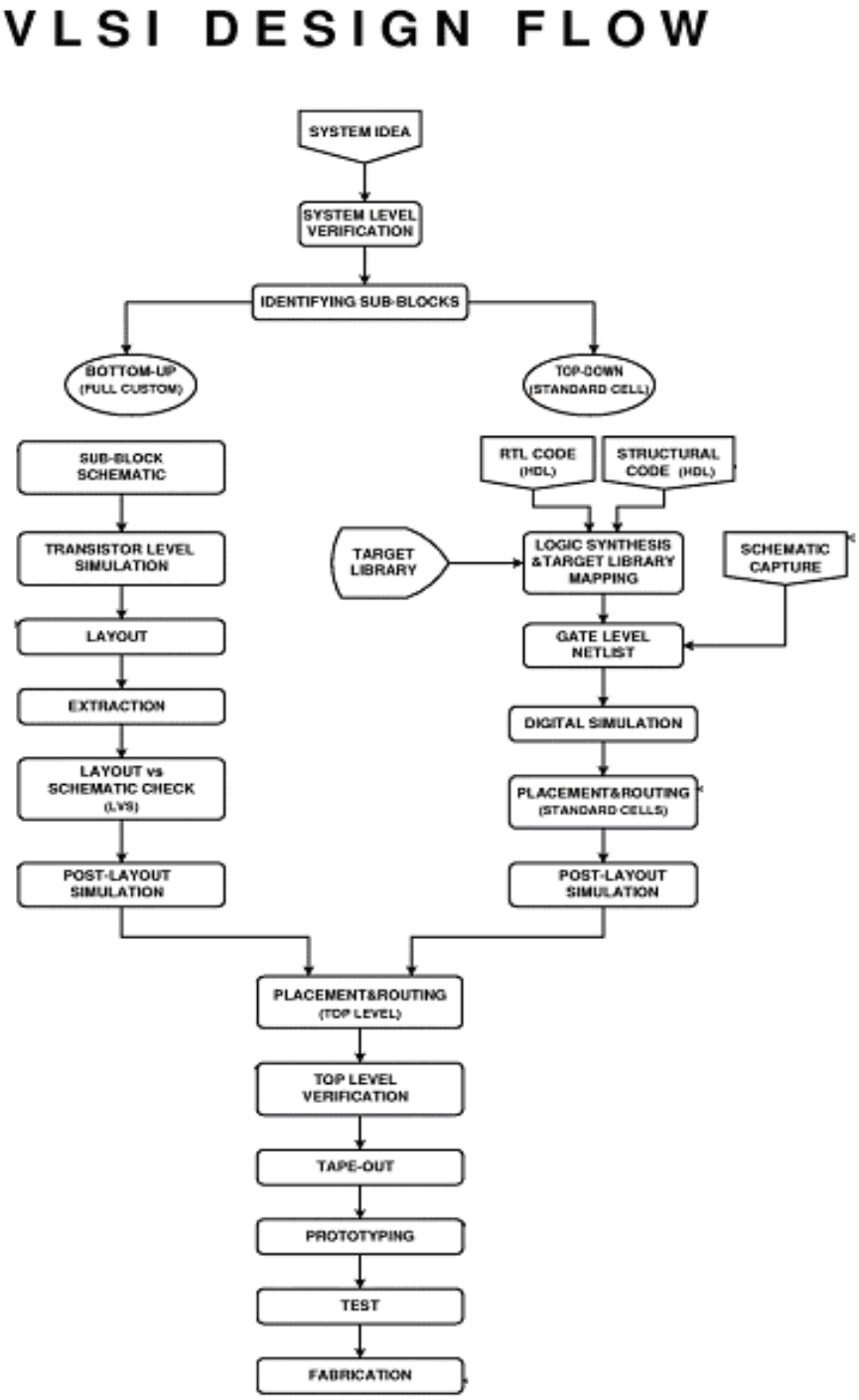
- VLSI Design Flow
- MOS Layers
- Stick Diagrams
- Design Rules and Layout
- 2 □m CMOS Design rules for wires
- Contacts and Transistors Layout Diagrams for NMOS and CMOS Inverters and Gates
- Scaling of MOS circuits
- Limitations of Scaling

UNIT 2

INTRODUCTION:

Stick diagrams and layout diagrams are used for the representation of different components and circuits in **VLSI**. All these aspects are explained in this chapter.

VLSI DESIGN FLOW



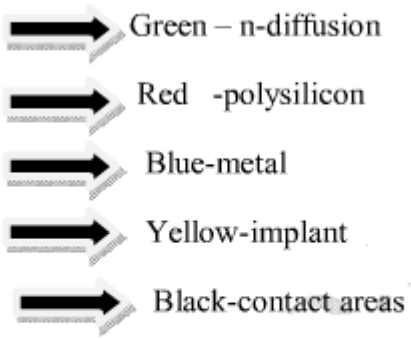
MOS LAYERS

MOS circuits are formed from four basic layers – n-diffusion, p-diffusion, polysilicon and metal. They are isolated from each other with thinox (thin silicon dioxide) layers. When polysilicon and thinox regions cross one another a transistor is formed. Layers are joined together to form contacts.

The actual circuit in silicon is represented by simple diagrams which convey both layer information and topology.

STICK DIAGRAM

Stick diagrams are used to convey layer information though the use of color code. In the case of nMOS design:-



Monochrome encoding of the lines is used to represent the diagrams in black and white.

The simple set required for a single metal nMOS design is set out in first figure.

For double metal cmos p-well process, encodings are in second figure.

N-MOS DESIGN STYLE:-

The layout of nMOS involves the color encoding given previously and a transistor is formed wherever polysilicon crosses n-diffusion (red over green)

All diffusion wires (inter connections) are n-type (green)

➡ First step is to draw the metal (blue) V_{DD} and GND rails with enough space between them

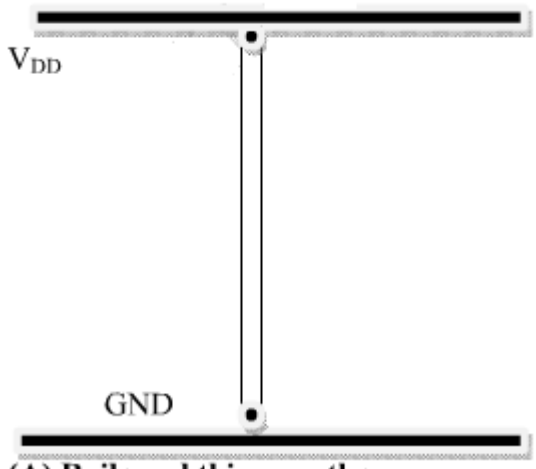
➡ Thinox (green) paths may be drawn between the rails for inverters.

➡ A depletion mode transistor is connected between output point to V_{DD} and a pull down structure of enhancement mode are suitably connected between output point and ground.

➡ Long signal paths require metal buses (blue)

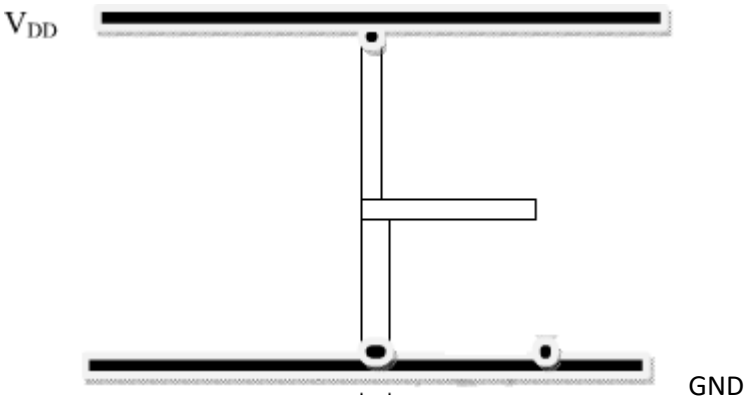
Examples of nMOS stick layout design style.

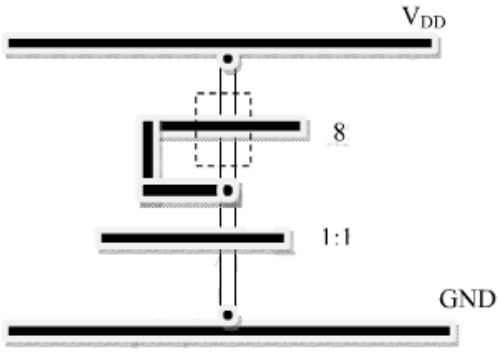
REGISTER CELL

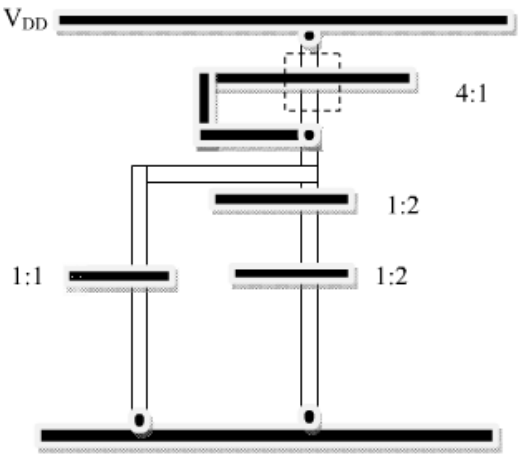


(A) Rails and thinox paths

(2) LOGIC FUNCTION X=A+B.C





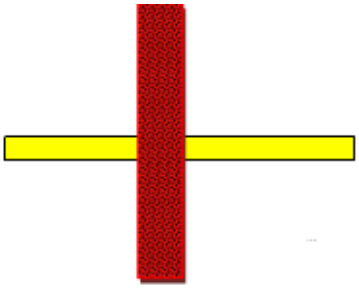


CMOS design style:

CMOS representations are extension of NMOS approach.

Yellow in cmos design is used to identify p-transistors and wires, as depletion mode devices are not used.

N-type (red over green)



P-type (red over yellow)

⇒ First the V_{DD} and V_{SS} rails are drawn and then demarcation line is drawn

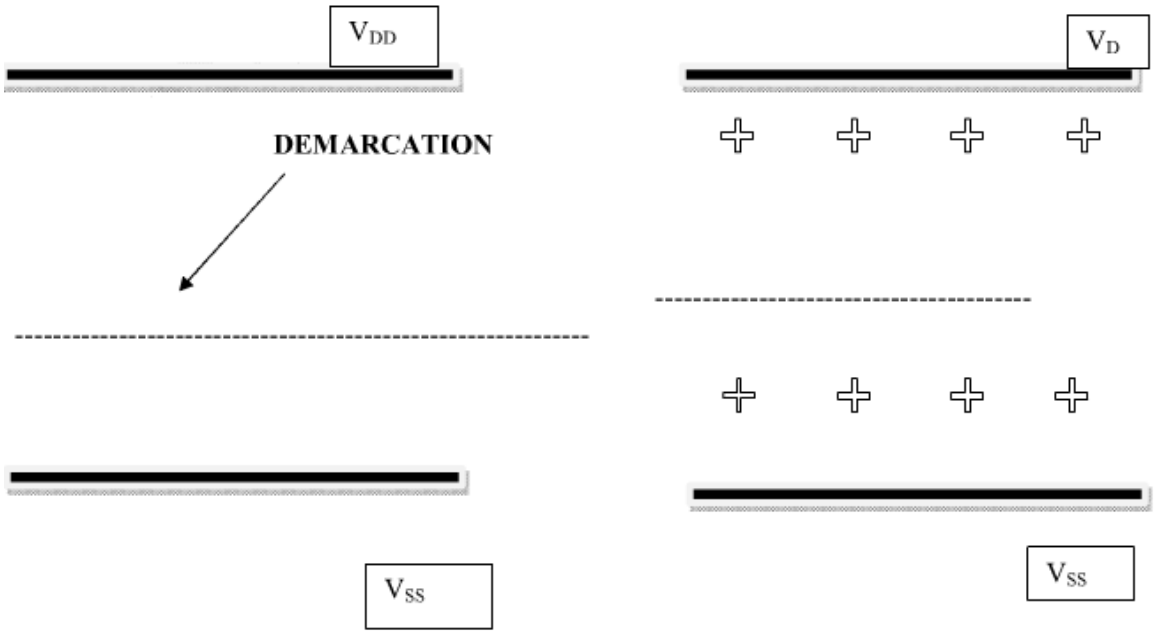
⇒ P and n transistors are drawn

⇒ Metal and diffusion connections are made

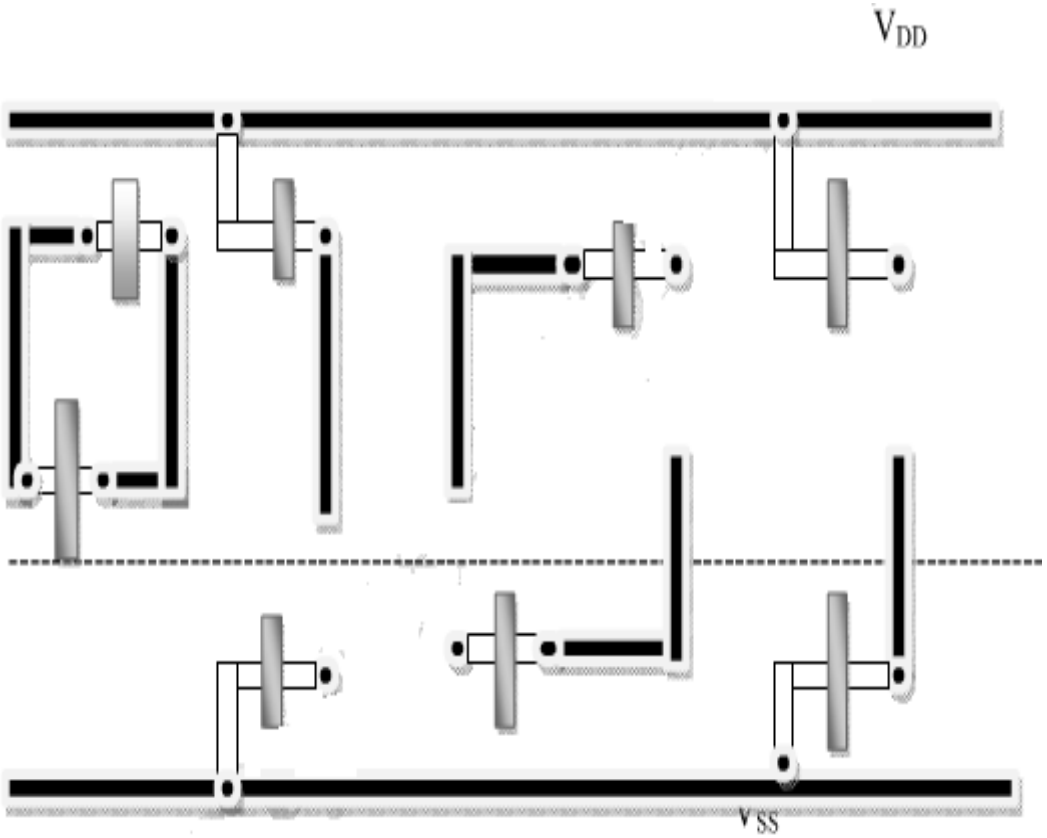
⇒ Remaining interconnections are made

Examples of cmos stick layout design style.

(Using a 1-bit shift register stage as an example)

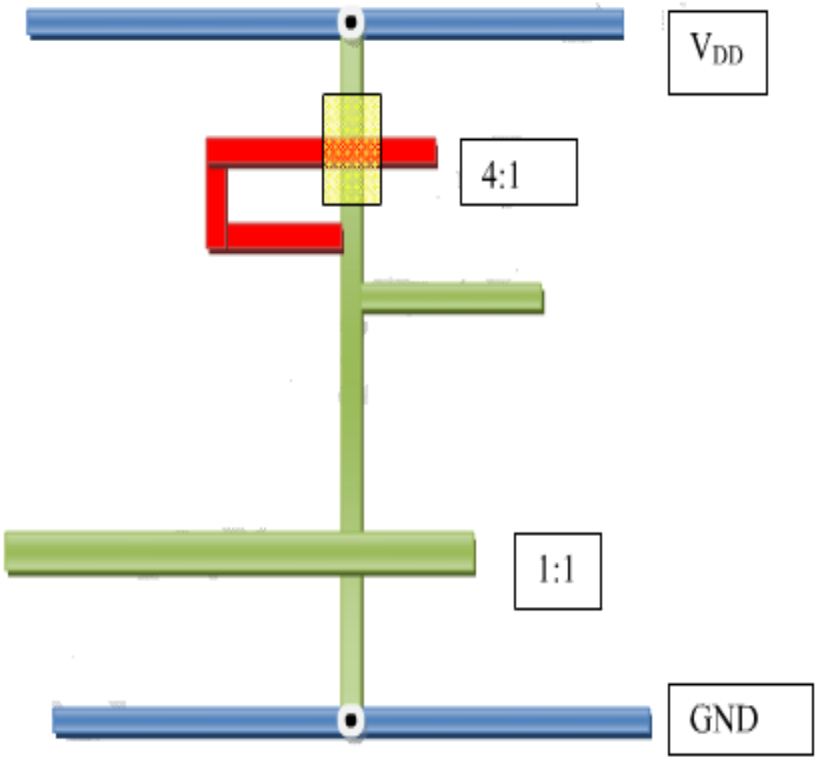


(A)RAIL AND DEMARCATION LINE TRANSISTORS



METAL AND DIFFUSION CONNECTIONS

stick diagram for n-mos inverter



NMOS INVERTER

DESIGN RULES AND LAYOUT

The aim of design rules is to allow a ready translation of circuit design concepts, usually in stick diagram or symbolic form into actual geometry in silicon.

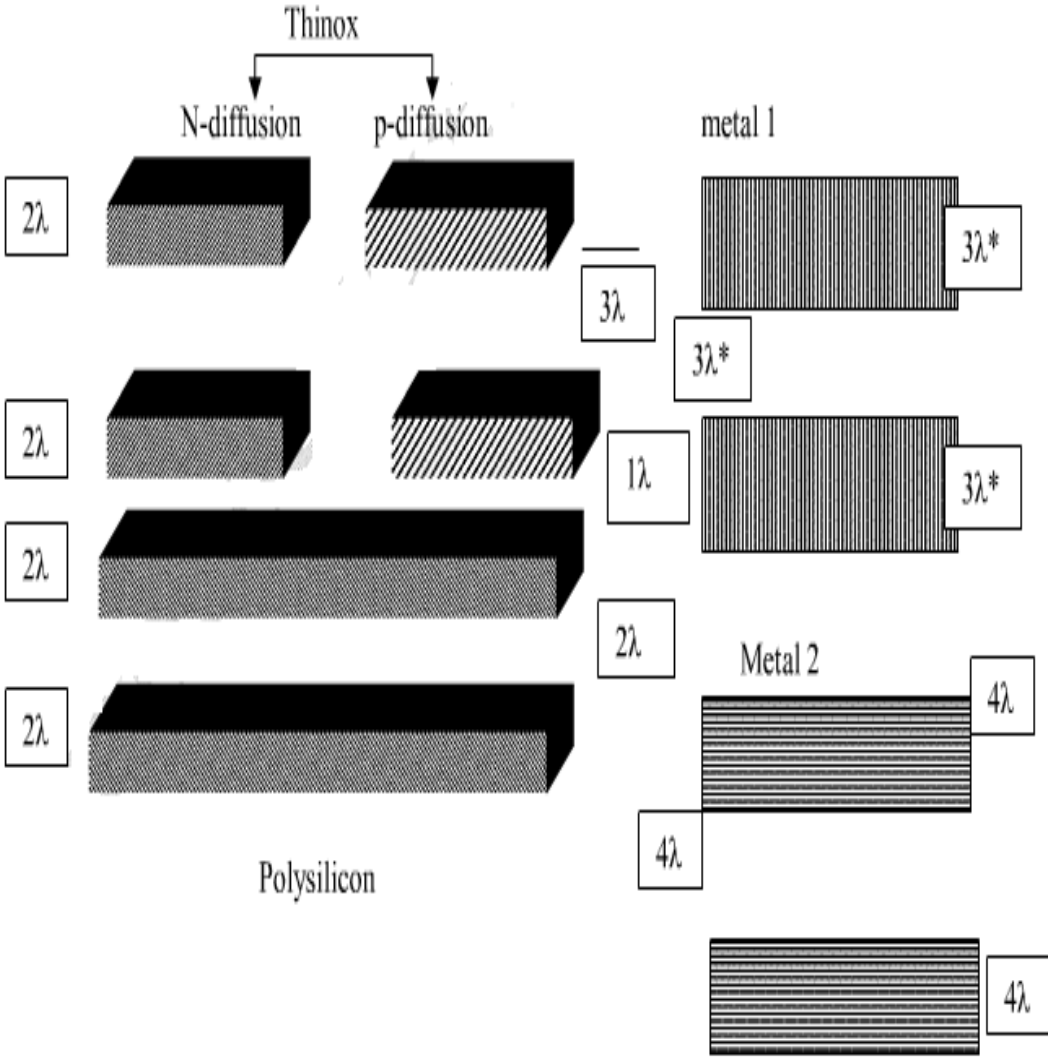
LAMBDA-BASED DESIGN RULES:-

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.

Design rules for wires (nMOS and CMOS)

Minimum width specified)

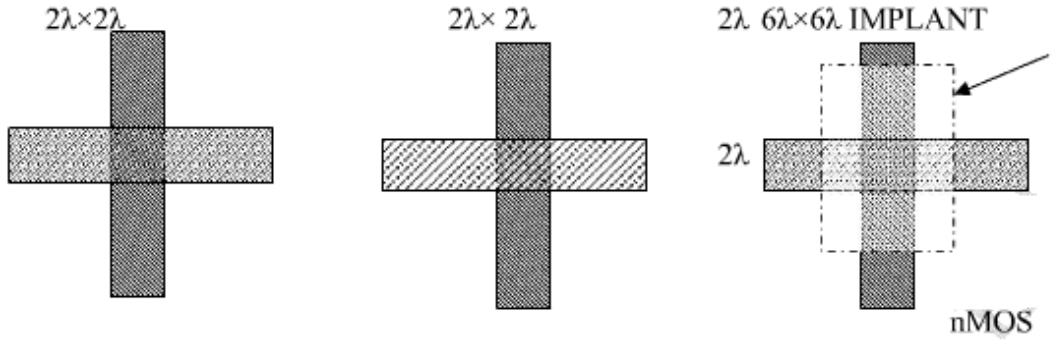
minimum separation (where



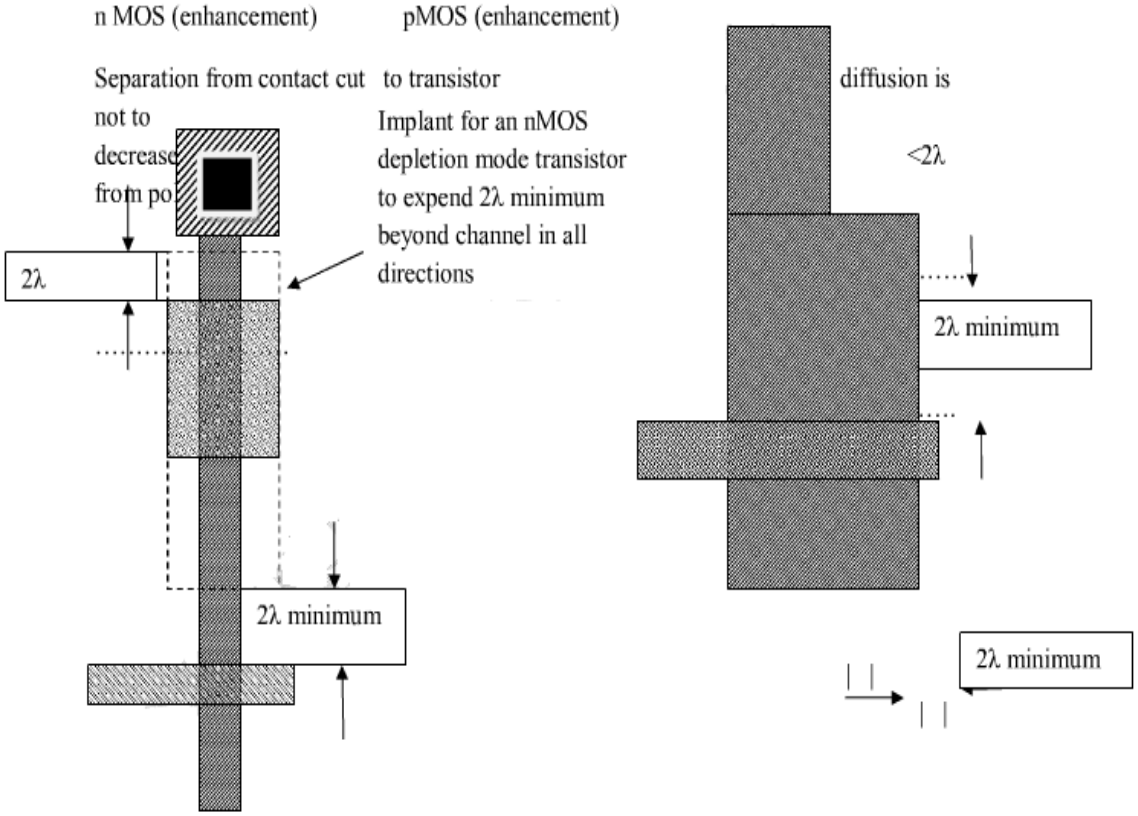
Where no separation is specified, wires may overlap or cross (e.g. metal is not constrained by any other layer) for p-well cmos note that n-diffusion wires can only exist inside and p-diffusion wires out side the p-well.

Note: many fabrication houses now accept 2λ metal 1 width and separation.

Transistor design rules (nMOS, pMOS mos and CMOS)



(depletion)



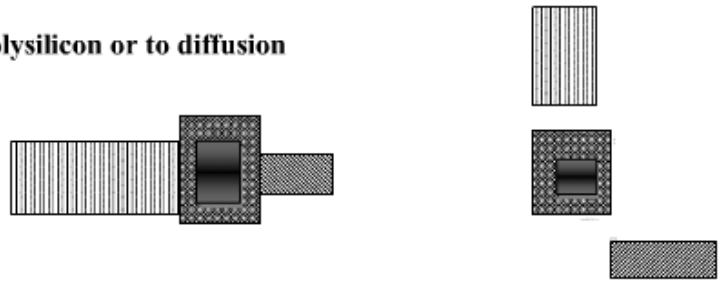
polysilicon to extend a minimum of 2λ beyond diffusion boundaries (width constant)

Separation from implant to another transistor

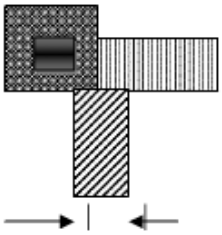
Key: polysilicon n-diffusion p-diffusion transistor channel

(Polysilicon over thinox)

metal 1 to polysilicon or to diffusion

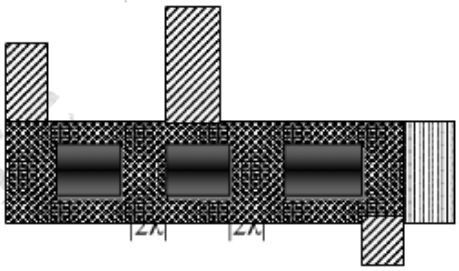


3λ minimum



cuts

2λ minimum



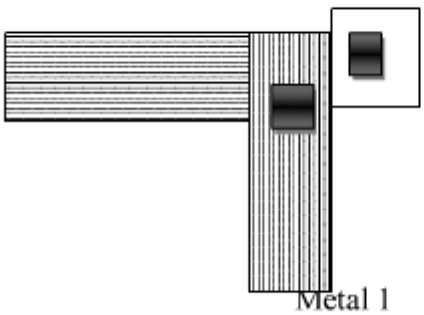
minimum separation multiple

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

Via (contact from metal 2 to metal1 and hence to other layers)

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

Metal 2

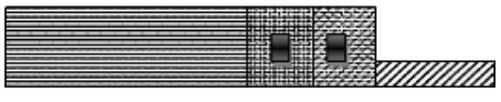


Metal 1

4λ×4λ area of overlap with

2λ×2λ via at center

Via and cut used
to connect metal 2
to diffusion

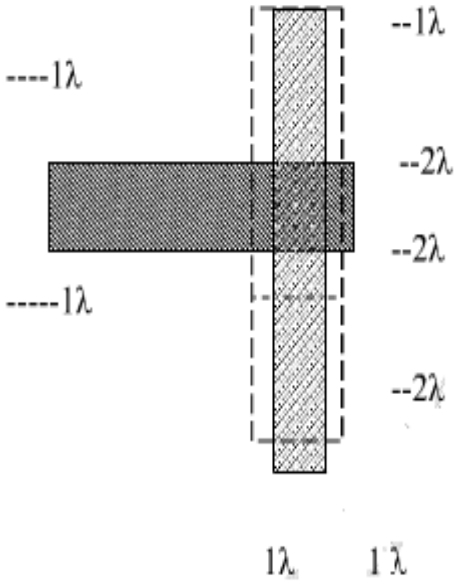


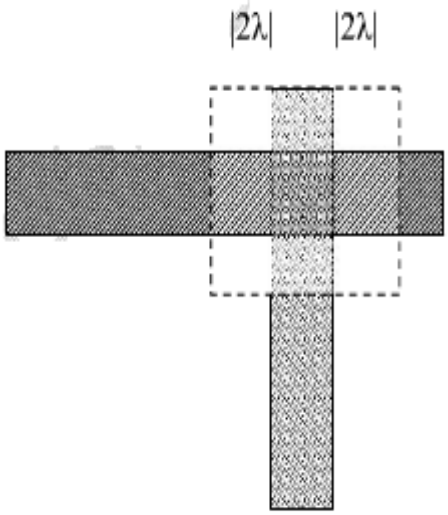
Via cut

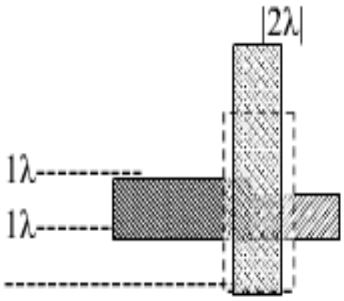
Contacts (nMOS and CMOS)

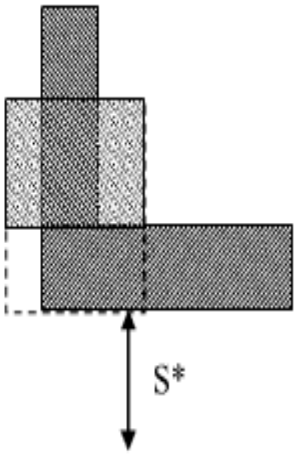
CONTACT CUTS

When making contacts between polysilicon and diffusion in nMOS circuits it should be recognized that there are three possible approaches poly to metal to diffusion or a buried contact poly to diffusion

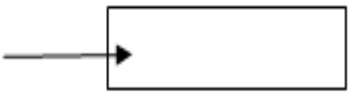






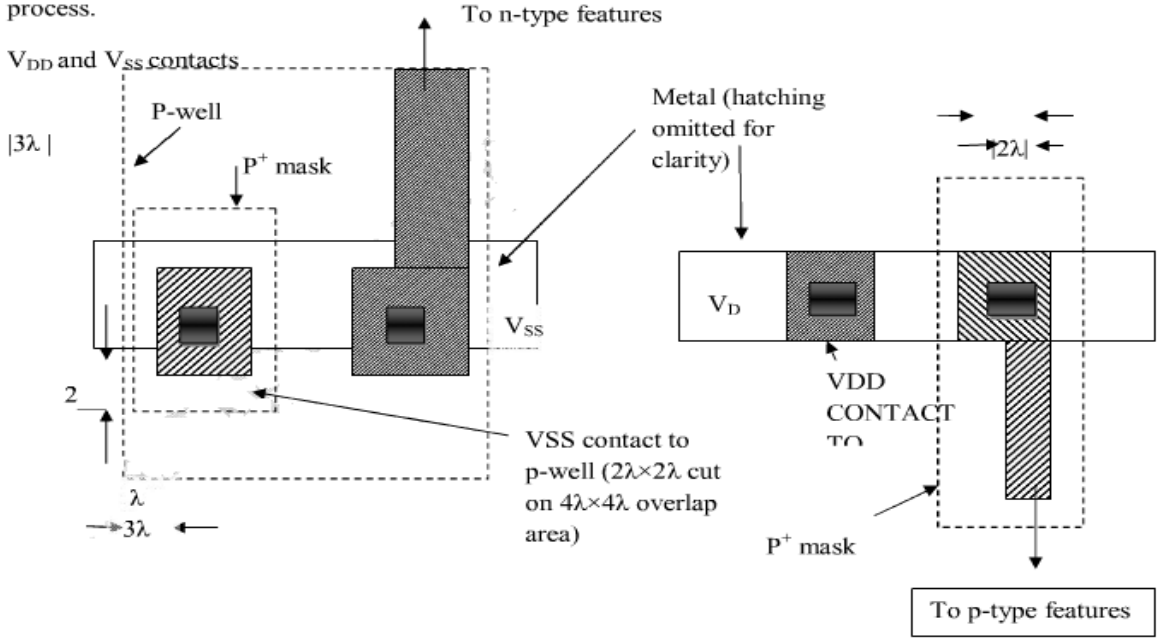


Unrelated polysilicon or diffusion

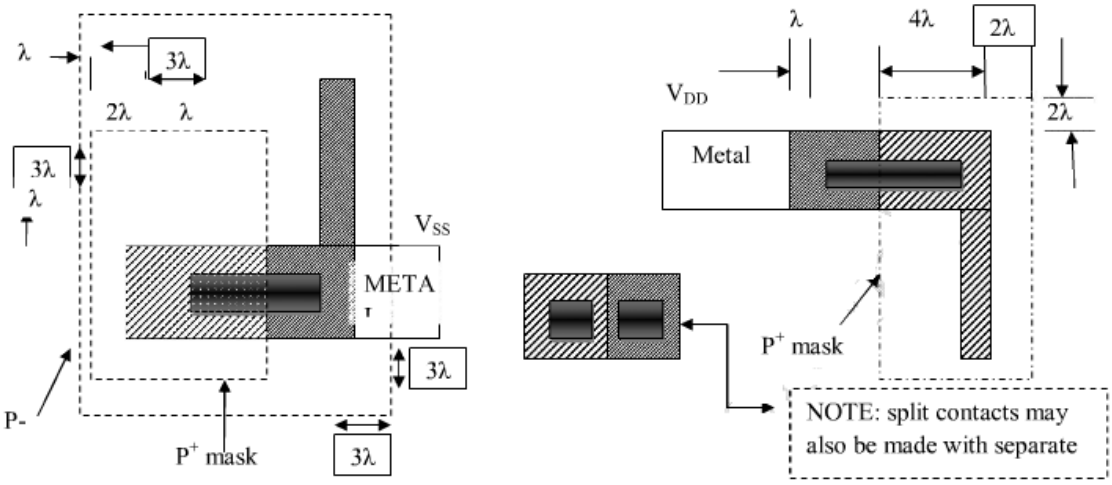


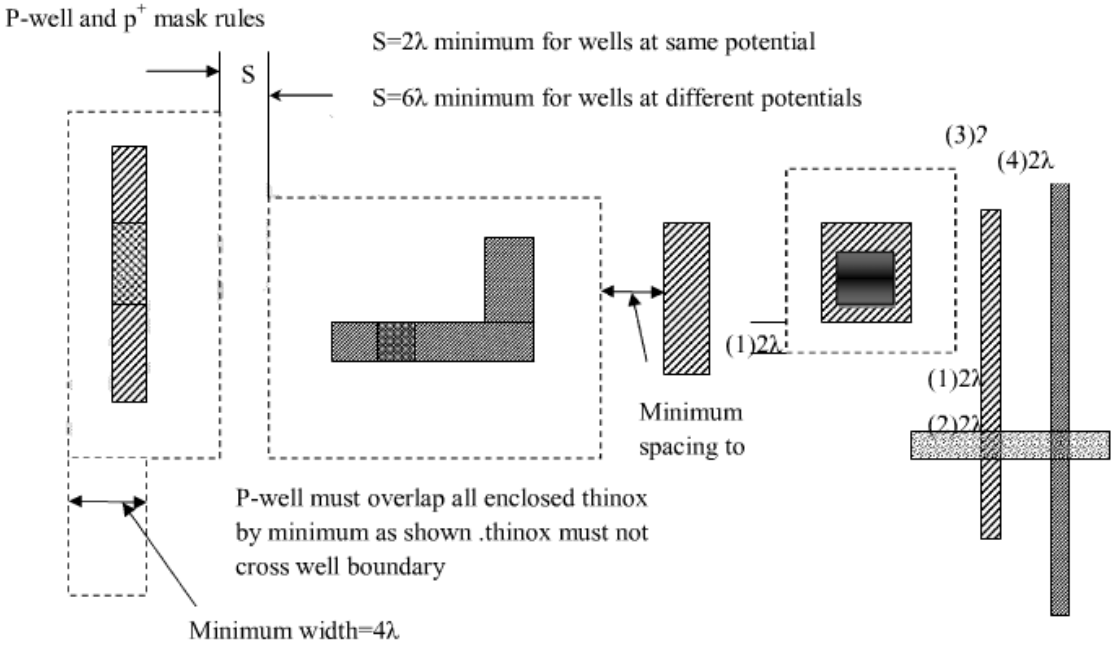
CMOS LAMDA BASED DESIGN RULES

Cmos fabrication is much more complex than n-mos fabrication. These rules form the abstract of the exact manufacturing process.



Each of the above arrangements can be merged into single 'split' contacts.



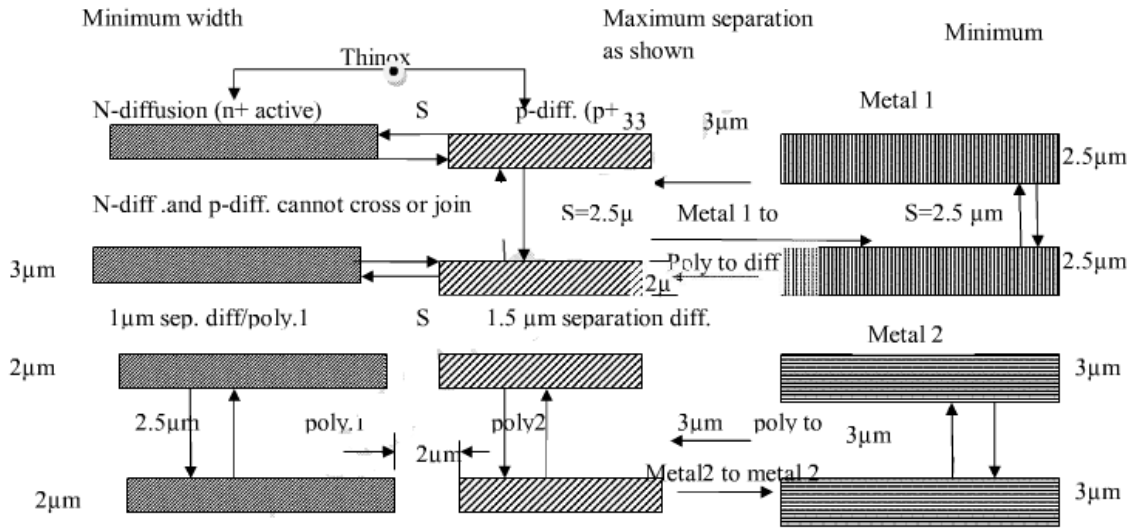


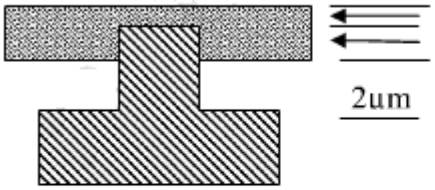
µM CMOS DESIGN RULES

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.



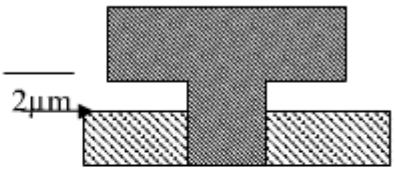


1.5 µm min.edge to edge

2µm

1.5 µm min. overlap

1.5µm min. overlap capacitors poly. 1/poly.2



Poly 1 Overlapping poly.2

Poly. 2 overlapping poly.1

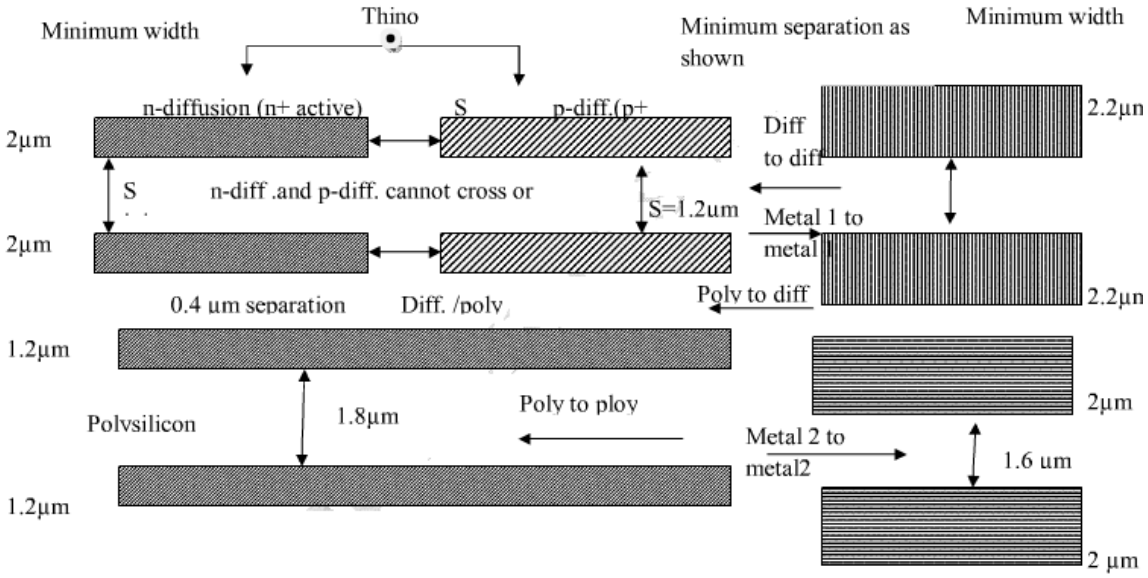
AVOID COINCIDENT EDGES WHERE METAL 1 AND METAL 2 RUNS FOLLOW THE SAME PATH FOR >25µm LENGTH (UNDER LAP METAL 1

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

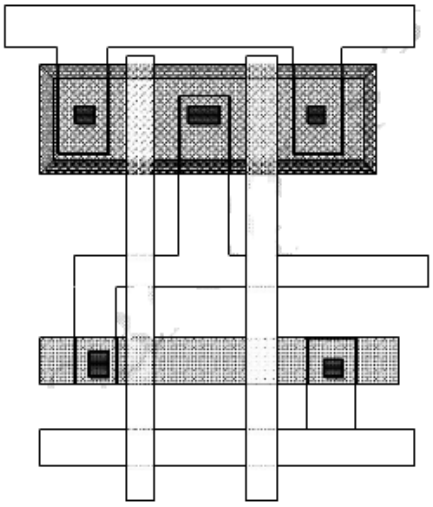
2µm DOUBLE METAL, SINGLE POLY CMOS RULES

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

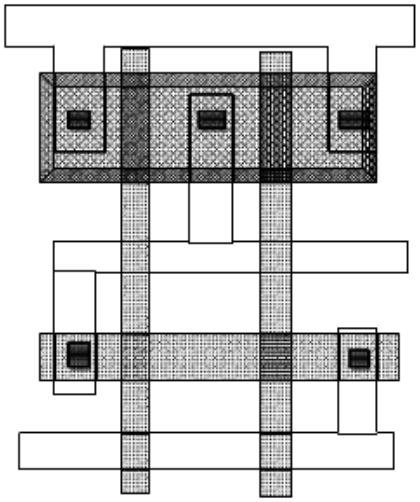
Design rules for wires (interconnects) (orbit 1.2 µm CMOS)



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).

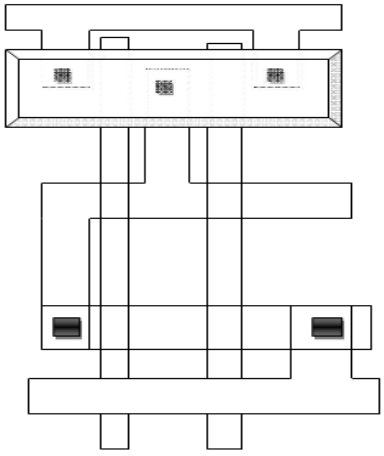


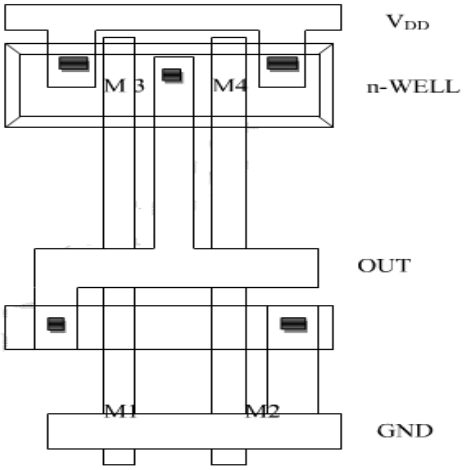
N-WELL AND ACTIVE AREA MASKs AND . ..



POLY MASK -> DEFINE NMOS

…….PMOS TRANSISTORS



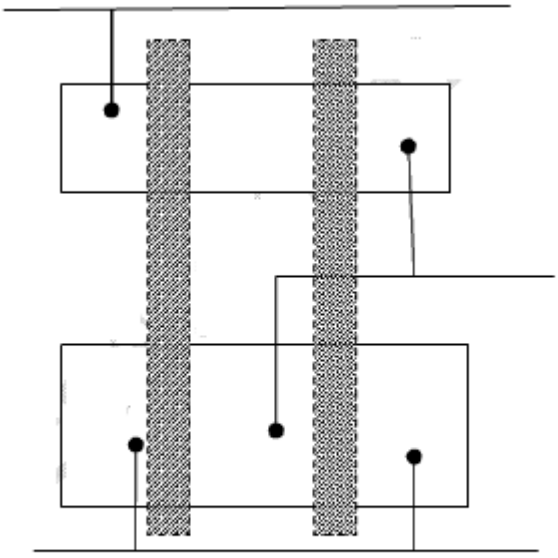


Metal mask for V_{DD}, GND and output connections

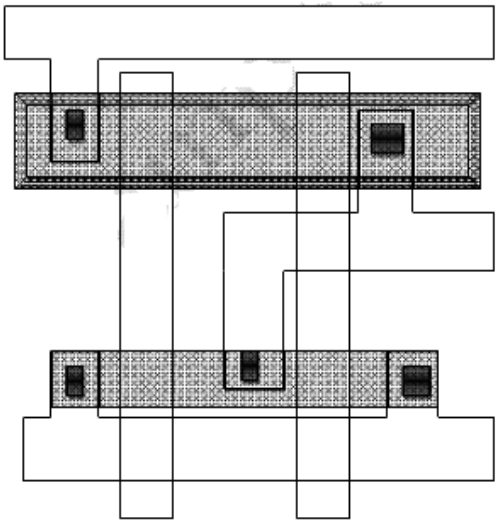
V_A V_B

METAL –DIFFUSION CONSTANT MASK

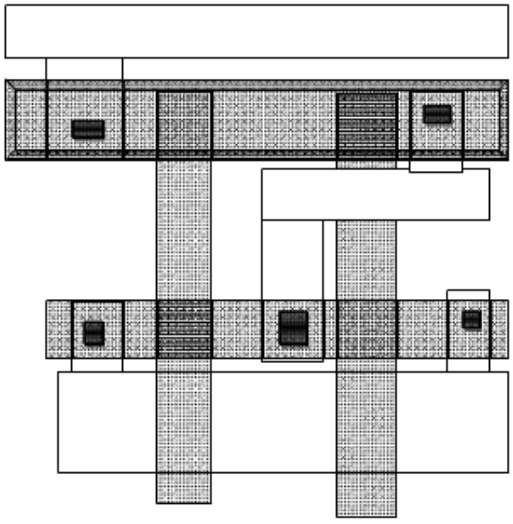
MASK LAYOUT OF A CMOSNOR GATE:



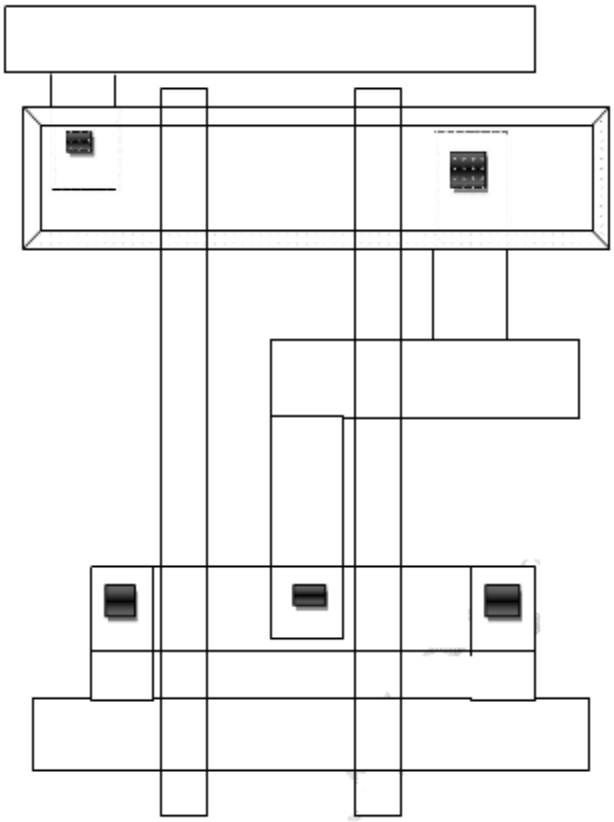
STICK DIAGRAM LAYOUT



N-WELL AND ACTIVE AREA MASKS

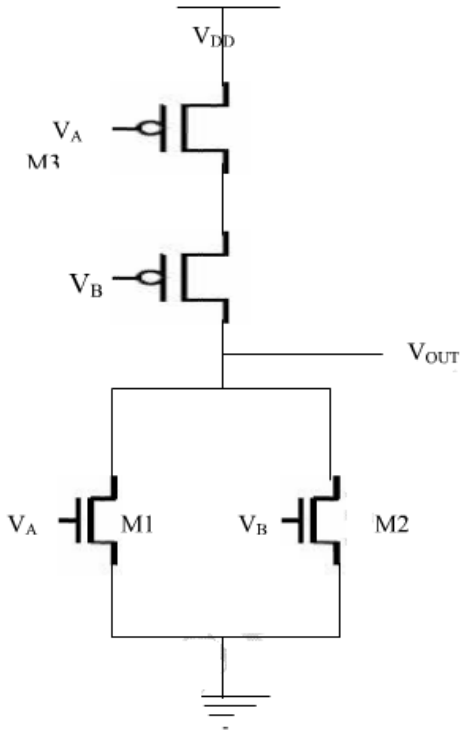


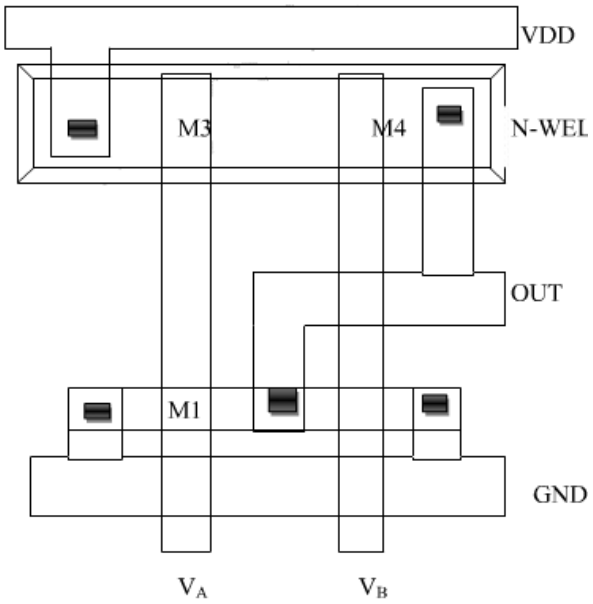
POLY MASK-> DEFINE NMOS AND PMOS TRANSISTORS



METAL MASK FOR V_{DD}, GND AND OUTPUT CONNECTIONS

TRANSISTOR AND STICK DIAGRAM REPRESENTATION :



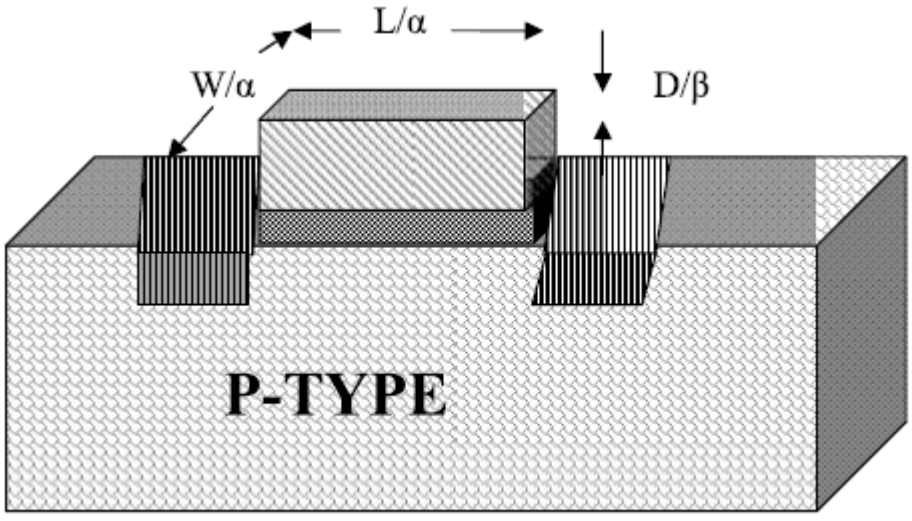


SCALING OF MOS CIRCUITS

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

SCALING FACTORS FOR DEVICE PARAMETERS:

GATE AREA AG :

$$\mathbf{Ag = L.W}$$

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

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

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

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 $1/\beta$

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

GATE CAPACITANCE C_g:

$$\mathbf{C_g = C_O\ L.W}$$

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}

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

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

Thus Q_{on} is scaled by 1.

CHANNEL RESISTANCE R_{on}

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

Where μ 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 β^2/α^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_β is increased by $\ln \alpha$ 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 $\boldsymbol{\tau} = \mathbf{L/V_{drift}} = \mathbf{2d/\mu E}$

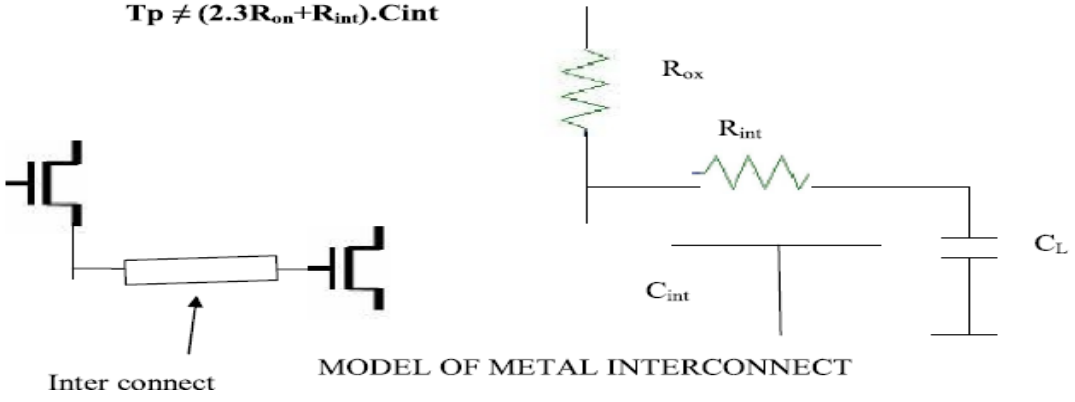
Limits due to interconnect and contact resistance

Since the width, thickness and spacing are scaled by $1/\alpha$, cross-section area must be scaled by $1/\alpha^2$ .Thus R is increased by α and I is scaled by $1/\alpha$. 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

$$\mathbf{Tp = R_{int}C_{int} + 2.3(R_{on}C_{int} + R_{on}C_L + R_{int}C_L)}$$

$$\mathbf{Tp \neq (2.3R_{on} + R_{int}).Cint}$$



MODEL OF METAL INTERCONNECT

Now

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

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

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.

ρ is the resistivity of interconnect L,W,H are the length, width and height of the interconnect.

UNIT-III

GATE LEVEL DESIGN

Contents:

- Logic Gates and Other complex gates
- Switch logic
- Alternate gate circuits
- Time delays
- Driving large capacitive loads
- Wiring capacitance
- Fan-in, Fan-out
- Choice of layers.

UNIT-3

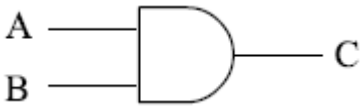
LOGIC GATES AND OTHER COMPLEX GATES

1) OR Gate:-

A	B	C
0	0	0
0	1	1
1	0	1
1	1	1

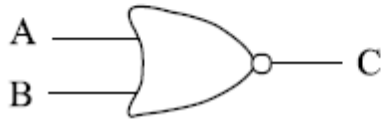
2) AND Gate:-

A	B	C
0	0	0
0	1	0
1	0	0
1	1	1



3) NOR Gate:-

A	B	C
0	0	1
0	1	0
1	0	0
1	1	0



4) NAND Gate:-

A	B	C
0	0	1
0	1	1
1	0	1
1	1	0

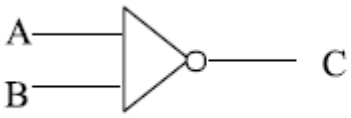
5) EX-OR Gate:-

A	B	C
0	0	0
0	1	1
1	0	1
1	1	0

6) EX-NOR Gate:-

A	B	C
0	0	1
0	1	0
1	0	0
1	1	1

7) NOT Gate:-

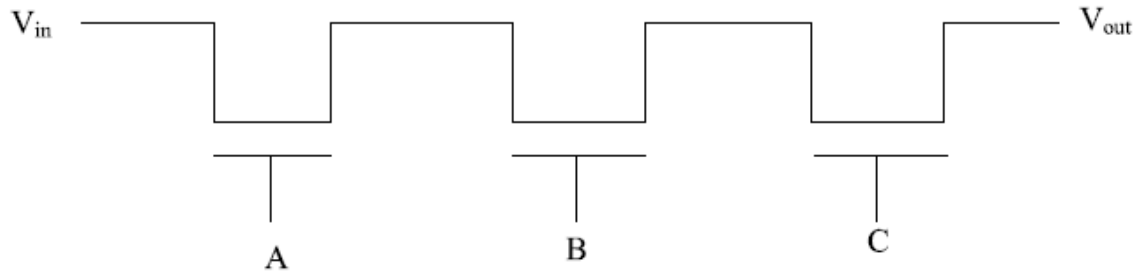


A	B
0	1
1	0

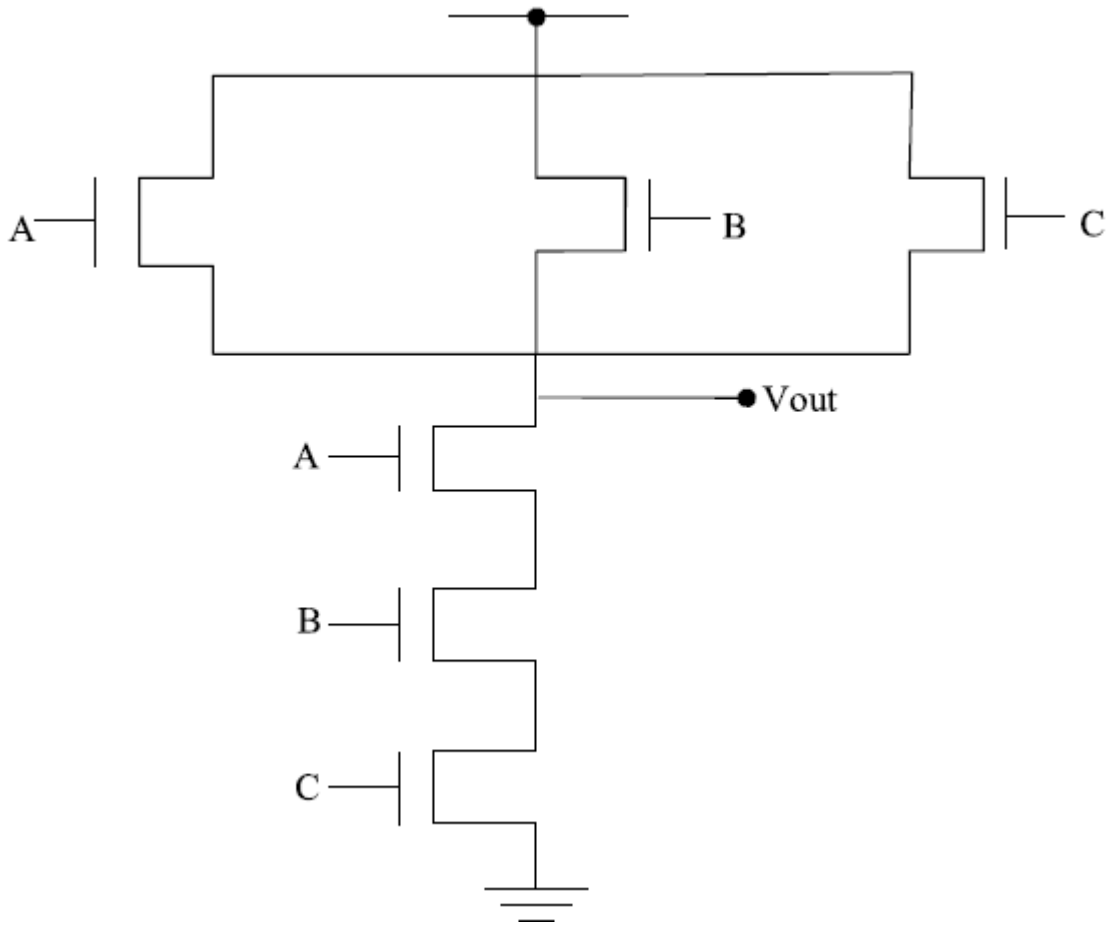
SWITCH LOGIC

It is based on the 'pass transistor' or on transmission gates. This approach is fast for small arrays. It is similar to logic arrays.

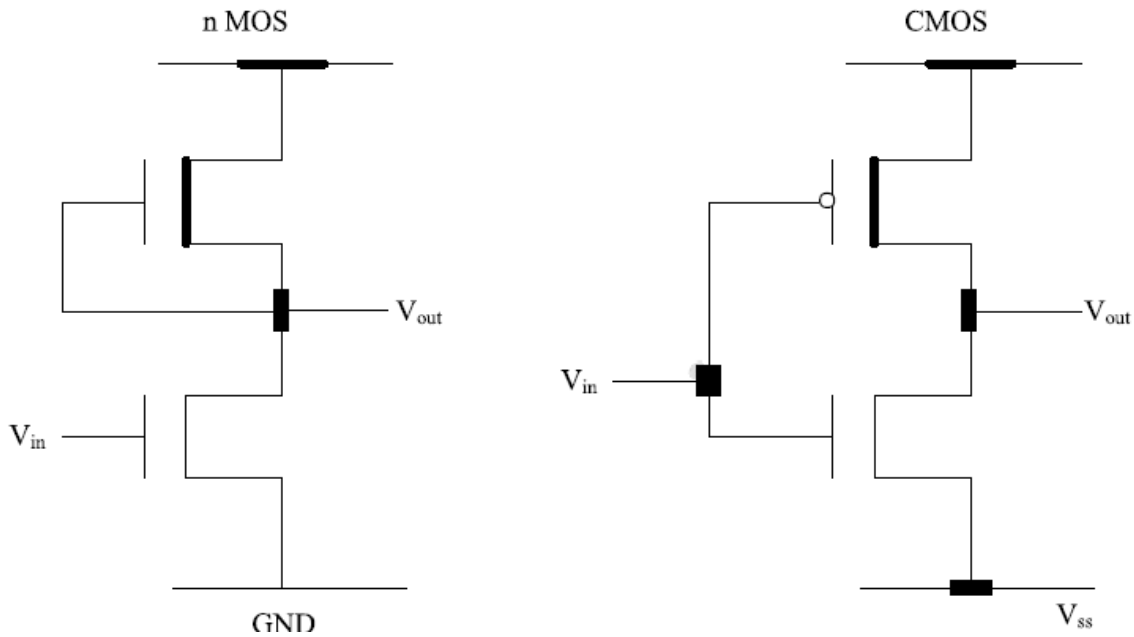
Many combinations of switches are possible as shown below:-



$V_{out} = V_{in}$ When ABC = 1

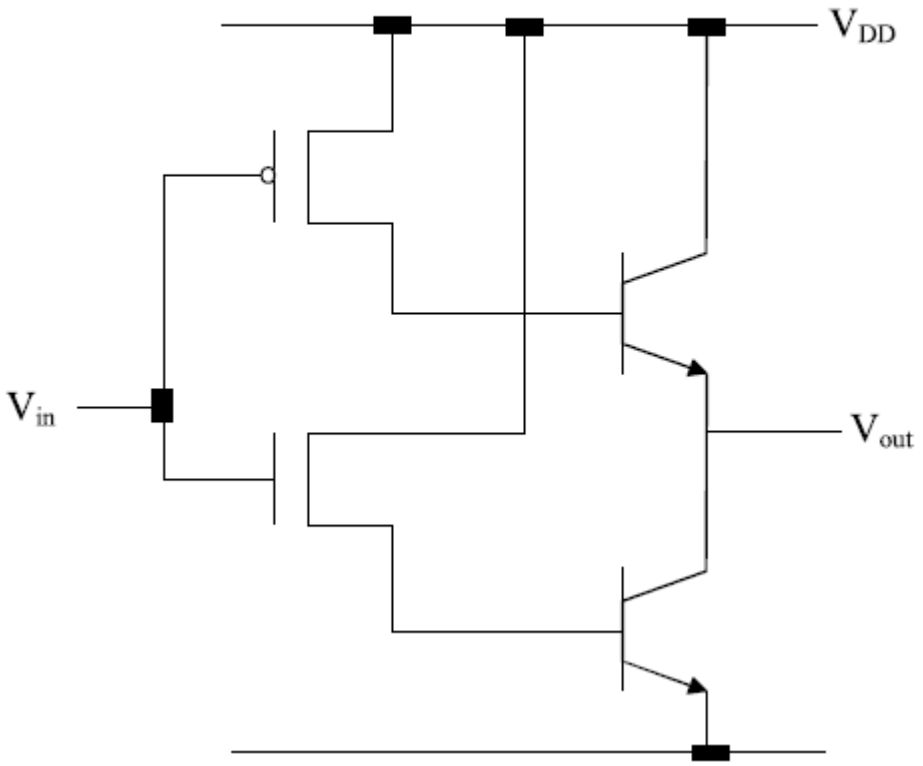


ALTERNATE GATE LOGIC



In gate logic the inverter is the simplest gate used. Both NAND and NOR and with CMOS, AND and OR gate arrangements are available. Some of the inverters used in gate logic are shown below:

Bi CMOS



SHEET RESISTANCE R_S

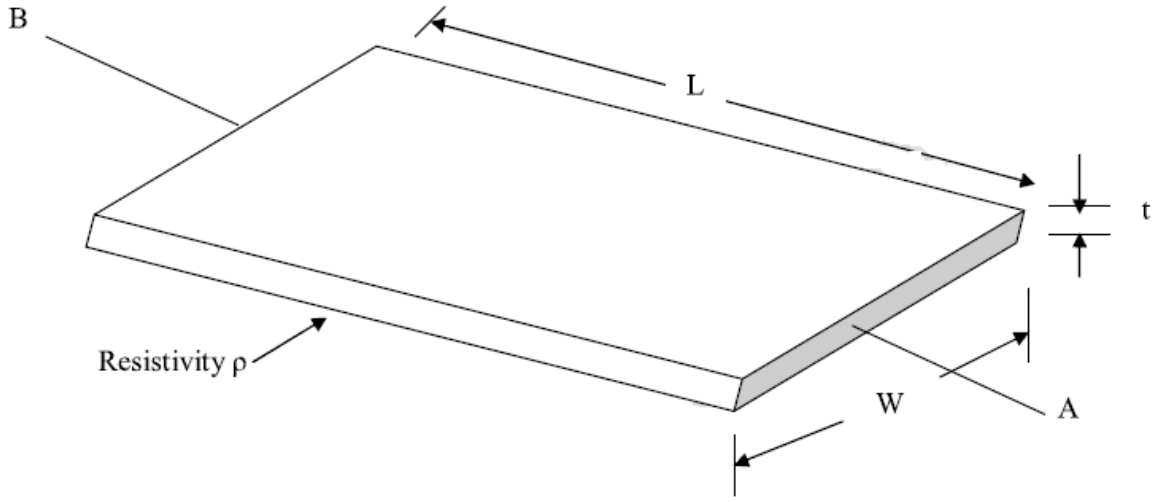
Consider a uniform slab of conducting material of resistivity ρ, of width W, thickness t, and length between faces L as shown below:

$$R_{AB} = \frac{\rho L}{tW} \quad \text{ohm}$$

Where A = cross section area.

Thus $R_{AB} = \frac{\rho L}{tW}$ ohm.

When L = W, i.e. a square resistive material, then



$$R_{AB} = \frac{\rho}{t} = R_s$$

Where R_s =ohm per square or sheet resistance.

Thus $R_s = \frac{\rho}{t}$ ohm per square.

It is completely independent of the area of the square.

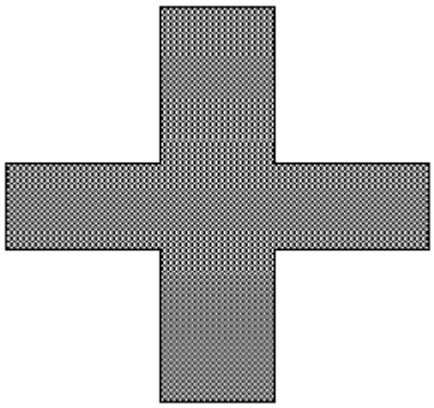
Typical sheet resistance R_s of MOS layers

Layer	R_s ohm per square		
	5μm	Orbit	1.2μm
Metal	0.03	0.04	0.04
Diffusion	10 → 50	20 → 45	20 → 45
Silicide	2 → 4	-	-
Polysilicon	15 → 100	15 → 30	15 → 30

n-transistor channel	10^4	2 X 10^4	2 X 10^4
p-transistor channel	2.5 X 10^4	4.5 X 10^4	4.5 X 10^4

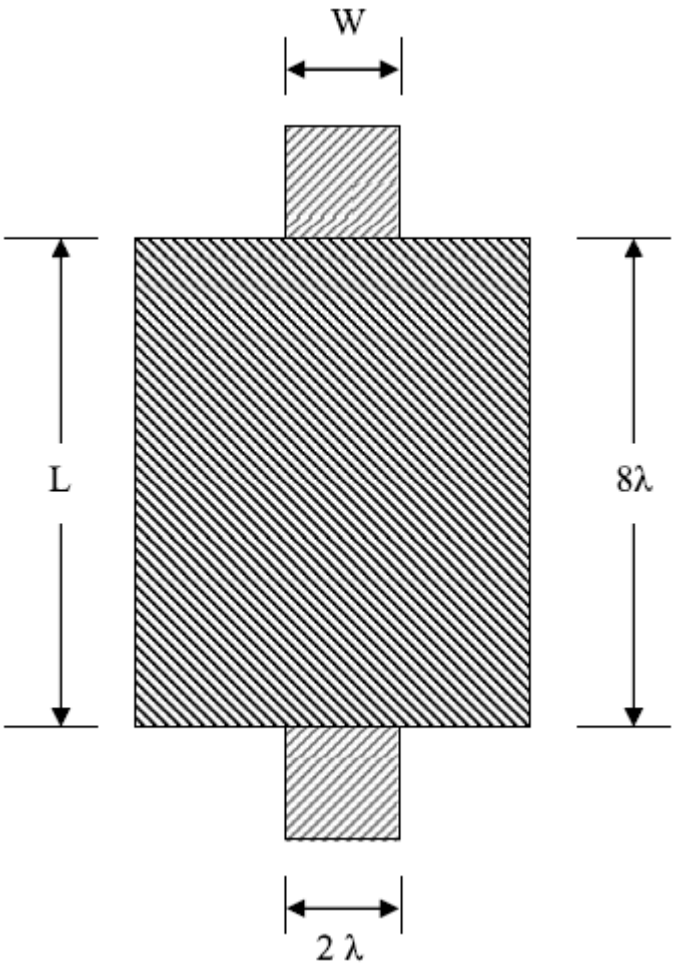
SHEET RESISTANCE CONCEPT APPLIED TO MOS TRANSISTORS AND INVERTERS

The simple n-type pass transistor has a channel length L = 2λ and a channel width W = 2 λ. The channel is square



$$R = \text{square} \times R_s \frac{Ohm}{square} = R_s = 10^4 \text{ ohm.}$$

The length to width ratio, denoted by Z is 1:1 in this case. Consider one more structure as in diagram below.



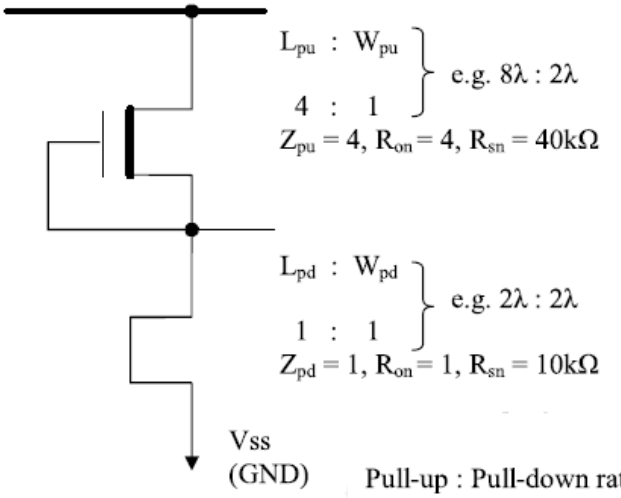
L = 8 λ and W = 2 λ

$$Z = \frac{L}{W} = 4$$

Channel resistance R = Z R_s = 4 X 10^4 Ohm.

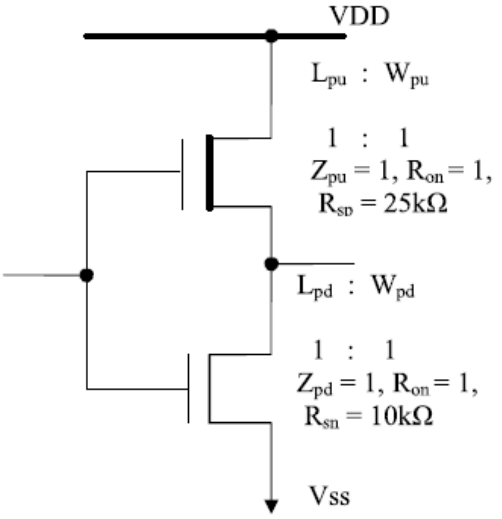
This channel can be taken as four 2 λ X 2 λ squares in series.

n MOS



Pull-up : Pull-down ratio = $Z_{pu} : Z_{pu}$ = 4 : 1 in this case.
On resistance (V_{DD} to GND) = 50kΩ

CMOS



A ratio rule does not apply and there is no static resistance from V_{DD} to V_{SS}.

AREA CAPACITANCES

Area capacitance can be calculated as $C = \frac{\varepsilon_o \varepsilon_{ins} A}{D}$ farads

Where

D = Thickness of silicon dioxide

A = Area of plates

ε_{ins} = Relative permittivity of SiO_2 = 4.0

ε_o = 8.85 X 10^{-14} F/cm (permittivity of free space)

The layer area capacitance is in pF/μm^2 (where μm = micron = 10^{-6} meter)

TABLE 4.2 Typical area capacitance values for MOS circuits

Capacitance	Value in pF X 10^{-4} / μm^2 (Relative values in brackets)		
	5 µm	2 µm	1.2 µm
Gate to channel	4 (1.0)	8 (1.0)	16 (1.0)
Diffusion (active)	1 (0.25)	1.75 (0.22)	3.75 (0.23)
Polysilicon to substrate	0.4 (0.1)	0.6 (0.075)	0.6 (0.038)
Metal 1 to substrate	0.3 (0.075)	0.33 (0.04)	0.33 (0.02)
Metal 2 to substrate	0.2 (0.05)	0.17 (0.02)	0.17 (0.01)
Metal 2 to metal 1	0.4 (0.1)	0.5 (0.06)	0.5 (0.03)
Metal 2 to polysilicon	0.3 (0.075)	0.3 (0.038)	0.3 (0.018)

Standard unit of capacitance C_g:-

A standard unit is employed that can be used in calculations. The unit is denoted as C_g and is defined as the gate-to-channel capacitance of a MOS transistor having W = L = feature size, that is a 'standard' or 'feature size' square.

C_g may be evaluated for any MOS process.

For example, for 5µm MOS circuits
Area/standard square = 5µm X 5µm = $25\mu m^2$

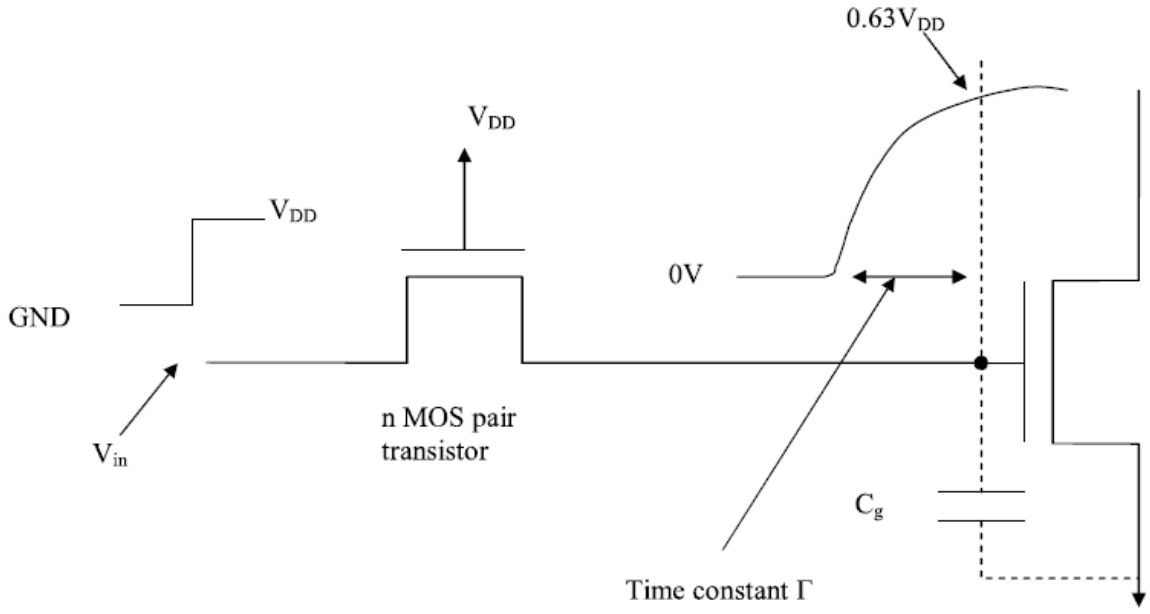
Capacitance value = 4 X 10^{-4} pF/μm^2
Thus standard value of C_g = 25 μm^2 X 4 X 10^{-4} pF/μm^2
= 0.01 pF

For 2 µm MOS circuits C_g = 0.0032 pF and for 1.2 µm MOS circuits C_g = 0.0023 pF

CALCULATIONS-T-DELAYS

The delay unit Γ is the product of 1 R_s and 1 C_g

$$\Gamma = (1\ R_s\ (\text{n-channel})\ X\ 1\ C_g)\ \text{seconds}$$



For 5μm technology

$\Gamma = 10^4$ ohm X 0.01 pF

= 0.1 n sec

For 2μm technology

$\Gamma = 2\ X\ 10^4$ ohm X 0.0032 pF

= 0.064 n sec

For 1.2μm (orbit) technology

$\Gamma = 2\ X\ 10^4$ ohm X 0.0023 pF

= 0.046 n sec

Practically Γ = 0.2 to 0.3 n sec for a 5μm technology because of circuit wiring and parasitic capacitances taken into account.

$$\tau \approx \tau_{sd} = \frac{L^2}{\mu_n V_{ds}} = \frac{25\ \mu m^2 V\ \text{sec}}{650\ cm^2 \quad 3V} X \frac{10^9\ n\,\text{sec}\ cm^2}{10^8\ \mu m^2}$$

= 0.13 n sec

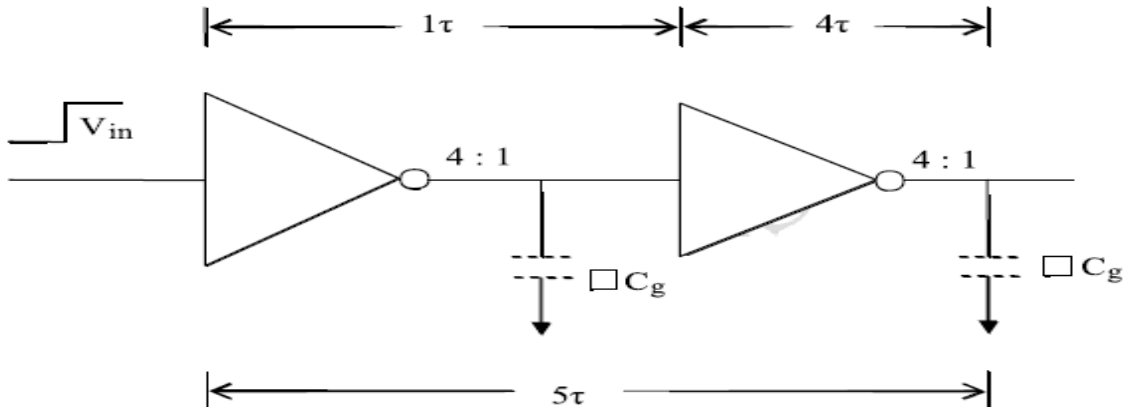
V_{ds} varies as C_g charges from 0 volts to 63% of V_{DD} in period Γ. Transit time and time constant Γ can be used inter changeably.

Inverter Delays:-

Consider 4 : 1 ratio nMOS inverter. To get 4 : 1 Z_{pu} to Z_{pd} ratio, R_{pu} will be 4 R_{pd}

$R_{pu} = 4\ R_s = 40k\Omega$
Meanwhile $R_{pd} = 1R_s = 10k\Omega$

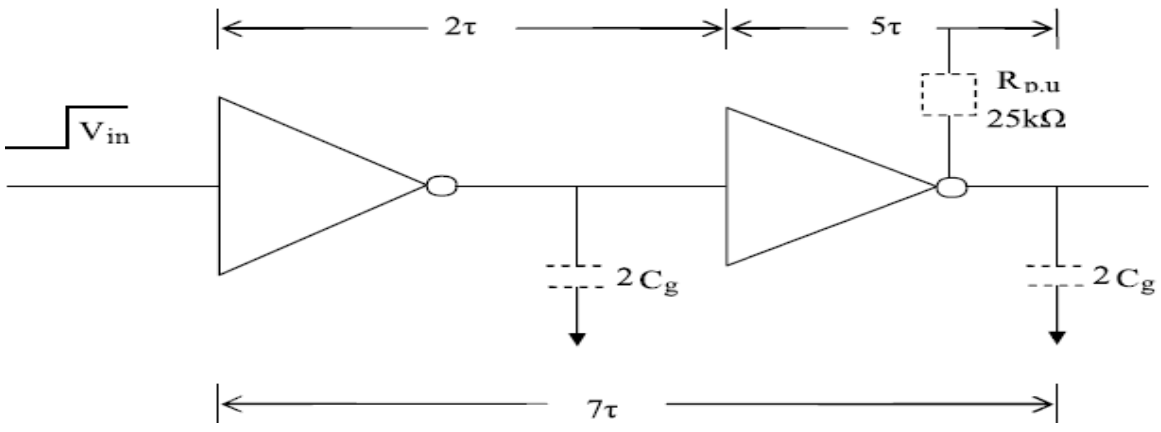
Consider a pair of cascaded inverters, the delay over the pair is constant. This is observed in diagram below:



Assuming $\tau = 0.3$ nsec, over all delay $= \tau + 4\tau = 5\tau$.

The general equation is $\tau_d = \left(1 + \frac{Z_{p.u}}{Z_{p.d}}\right)\tau$

Consider CMOS inverter, the nmos rule does not apply. The gate capacitance is $2\,C_g$ Because the input is connected to both transistor gates.



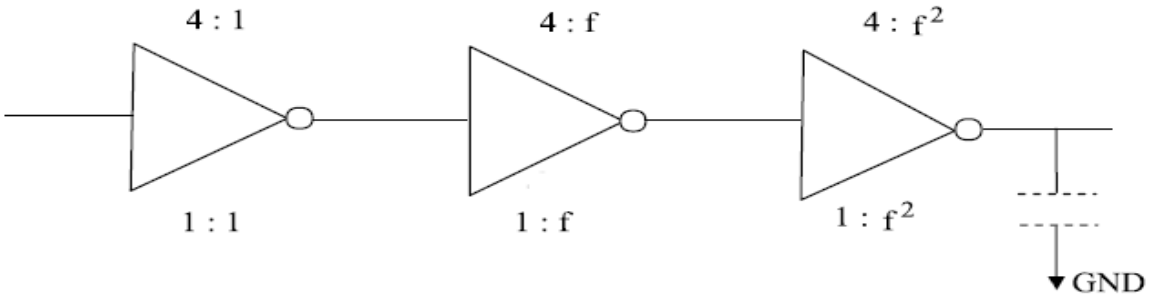
DRIVING LARGE CAPACITIVE LOADS

To drive large capacitive loads such as $C_L \geq 10^4 C_g$, they must be driven through low resistances, otherwise excessively long delays will occur.

CASCADED INVERTERS AS DRIVERS

Inverters intended to drive large capacitive loads must therefore present low pull-up and pull-down resistance. It implies L: W ratio must be low.

To eliminate this problem N cascaded inverters are used, each one of which is larger than preceding stage by a width factor f.

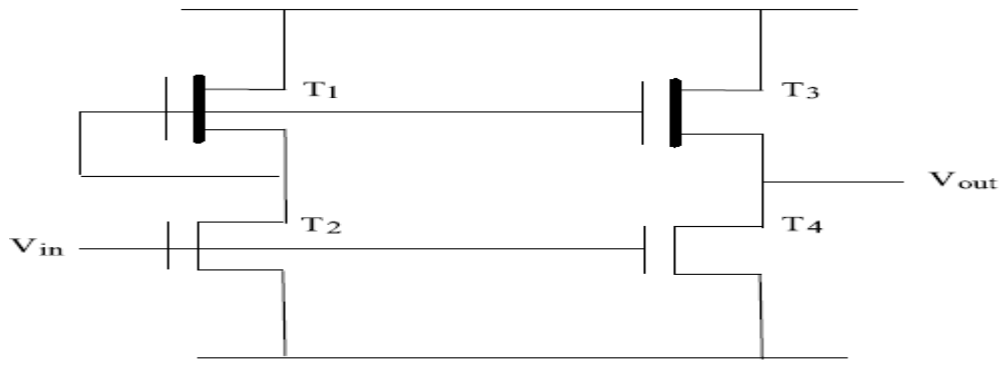


We take $y = \frac{C_L}{C_g} = f^N$

In all the cases, delay $\alpha\ Nf\tau = \frac{\ln(y)}{\ln(f)} f\tau$.

SUPER BUFFERS

A conventional inverter asymmetry gives rise to significant delay problems when it is used to drive more significant capacitive loads.



Inverting type n mos super buffer

WIRING CAPACITANCES

There are some sources of capacitances which contribute to the overall wiring capacitance. Three of them are discussed below:

1. Firing fields:

   It is the major component of the overall capacitance. C_{ff} is to be considered for accurate prediction of performance.

$$C_{ff} = \varepsilon_{si_{O_2}} * \varepsilon_0 * l \left[\frac{\Pi}{\ln\left(1 + \frac{2d}{t}\left(1 + \sqrt{1 + \frac{t}{d}}\right)\right)} - \frac{t}{4d} \right]$$

   Where, l = wire length
   t = thickness of wire
   d = wire to substrate separation
   Total wire capacitance $C_w = C_{area} + C_{ff}$

2. Inter layer capacitances:-

   It is highly dependant on layout. It occurs only where layers cross or when one layer underlies another.

3. Peripheral capacitance:-

   For n and p- regions formed by a diffusion process, the peripheral capacitance is important. For diffusion regions, each diode has a peripheral capacitance in Pico farads per unit length associated with it.

$$C_{total} = C_{area} + C_{periph}$$

FAN IN AND FAN OUT

FAN IN: - The number of similar gates that can be connected at the input is known as fan-in.

FAN OUT:-FAN OUT of a logic gate is defined as the number of inputs that the gate can drive without exceeding its worst case loading specifications.

It depends on the output characteristics.

It is of two types:

D.C fan out and A.C fan out.

D.C fan out:-The number of inputs that the output can drive with the output in a constant state (H or L).

A.C fan out:- The ability of an output to charge or discharge stray capacitance associated with the inputs that it drives.

The over all fan out will be the minimum of low state fan out and high fan out.

CHOICE OF LAYERS

The choice between the layers on which to route certain data and control signals should be considered.

- V_{DD} and V_{SS} should be distributed on metal layers.
- Long lengths of polysilicon should be less used.
- Capacitive effects should be carefully considered where fast signal lines are required.

The wires in a MOS system can be modeled as simple capacitors. Electrical rules for communication paths are established.

ELECTRICAL RULES

Layer	Maximum length of common wire		
	Lambda bared (5μm)	μm bared-(2 μm)	μm-bared (1.2 μm)
Metal	Chip wide	Chip wide	Chip wide
Silicide	2000λ	NA	NA
Polysilicon	200λ	400 μm	250 μm
Diffusion (active)	20λ	100 μm	60 μm

UNIT-IV

Contents:

- Subsystem Design
- Shifters
- Adders
- ALUs
- Multipliers
- Parity generators
- Comparators
- Zero/One Detectors
- Counters.
- SRAM,
- DRAM
- ROM
- Serial Access Memories
- Content Addressable Memory.

UNIT-4

SUBSYSTEM DESIGN:

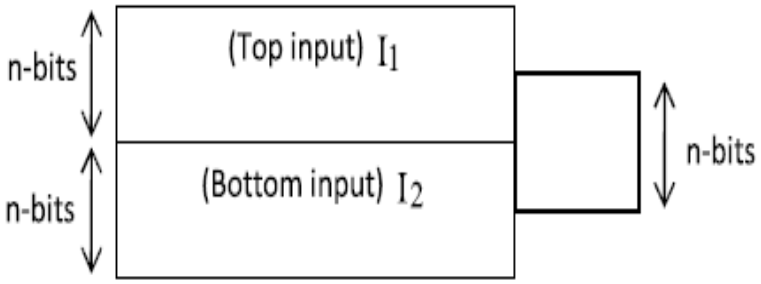
The guide lines for the subsystem design are as follows:

1. Define the requirements.
2. Partition the over all architecture into appropriate subsystem.
3. Consider communication paths carefully in order to develop sensible inter relationships between subsystems.
4. Draw a floor plan of how the system is to map onto the silicon (and alternate between 2, 3 and 4 as necessary).
5. Aim for regular structures so that design is largely a matter of replication.
6. Draw suitable (stick or symbolic) diagrams of the leaf cells of the subsystems.
7. Convert each cell to a layout.
8. Carefully carry out a design rule check on each cell.
9. Simulate the performance of each cell/subsystem.

SHIFTERS:

Barrel shifters:-

It can shift n-bits in a single combinational function. It also rotates and extends signs.



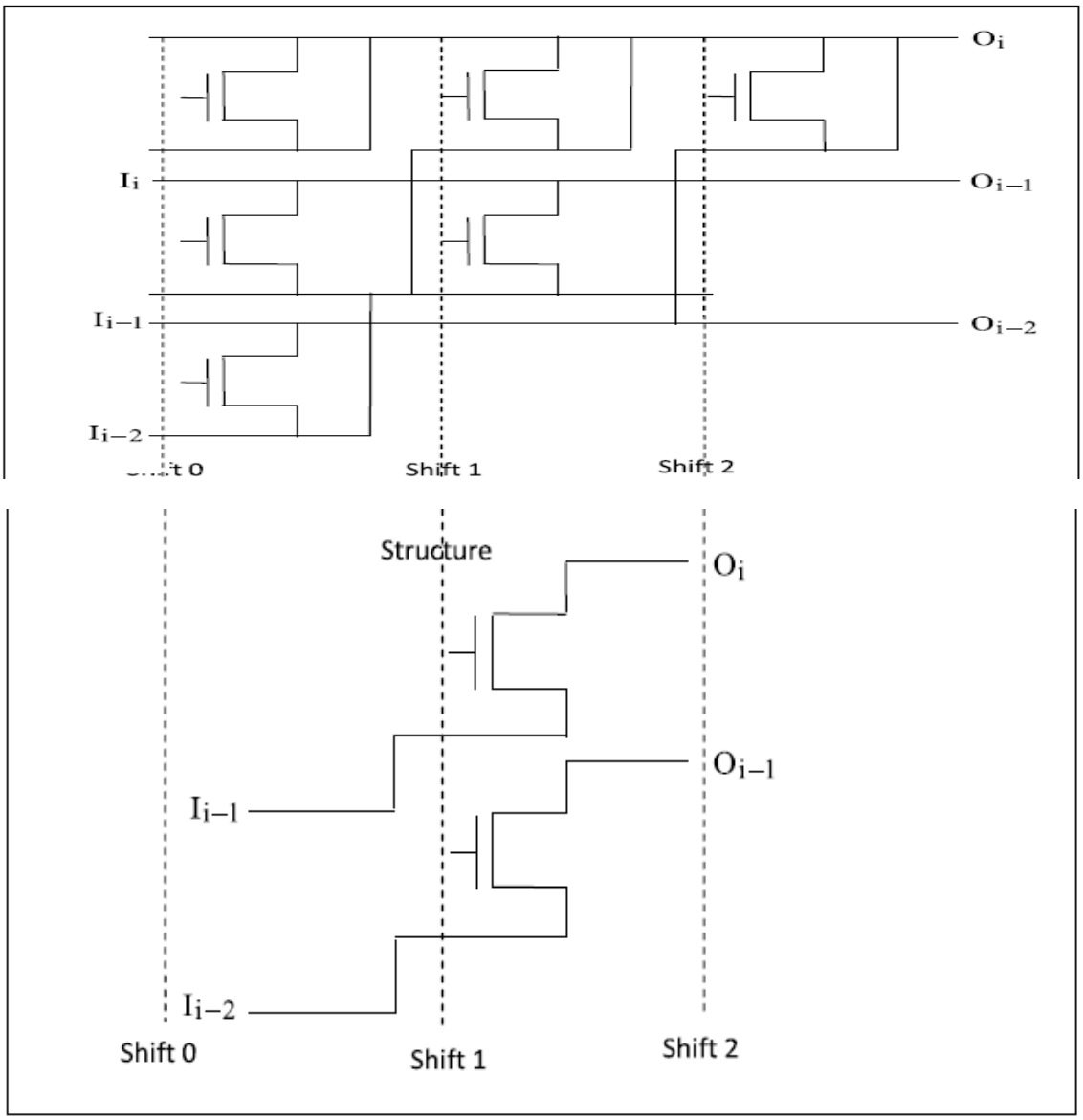
The above figure consists of 2n data bits and n control signals which can generate 'n' output bits and the shift operation may be performed by sending an n-bit slice of the 2n data (n+n) bits to the outputs. The actual operation is found by the values placed at the data inputs.

Example:-

Place a data word dw into the top input and word '0' (all zeroes) into the bottom input.

Now the output is right shift with zero. Here the control bits select the top most n bits by a shift of zero and select the bottom most n-bits by an n-bits shift that pares the word out of the shifter.

A circuit can be constructed by using barrel shifter, which has 'n'output bits from 2n vertical by n horizontal arrays of cells (each cell has 1-transistor and a few wires).



A section of barrel shifter

Let A be the input operand, B be the shift/rotate amount and Y be the shifted or rotated result. 'A' be an n-bit value, where n is an integer power of two. Therefore B is $\log_2(n)$-bit integer representing values from o to (n-1). Table below shoes all the operations performed by the barrel shifter for A = $a_7\ a_6\ a_5\ a_4\ a_3\ a_2\ a_1 a_0$ and B = 3 bits

Operation	Y
3-bit shift right logical	0 0 0 $a_7\ a_6\ a_5\ a_4\ a_3$
3-bit shift right arithematic	$a_7\ a_7\ a_7\ a_7\ a_6\ a_5\ a_4\ a_3$
3-bit rotate right	$a_2\ a_1\ a_0\ a_7\ a_6\ a_5\ a_4\ a_3$
3-bit shift left logical	$a_4\ a_3\ a_2\ a_1\ a_0$ 0 0 0
3-bit shift left arithematic	$a_4\ a_3\ a_2\ a_1\ a_0$ 0 0 0
3-bit rotate left	$a_4\ a_3\ a_2\ a_1\ a_0\ a_7\ a_6\ a_5$

ADDERS:

Addition is the most commonly used arithematic operation. Adder is most often the speed limiting element. So it should be carefully designed.

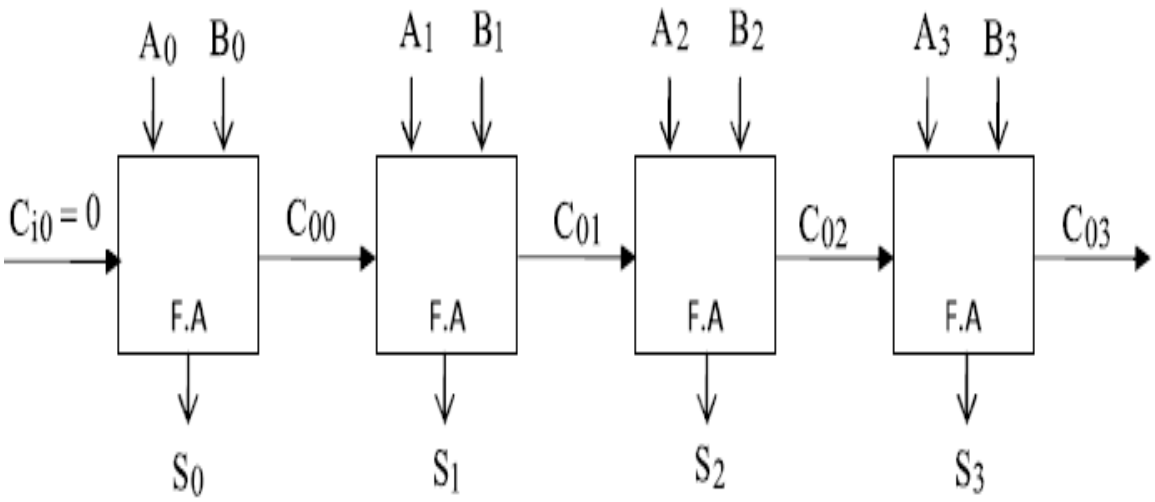
Binary Adder:-

The truth table for a binary full adder is given in the table below:-

A	B	C_i	S	C_o
0	0	0	0	0
0	0	1	1	0
0	1	0	1	0
0	1	1	0	1
1	0	0	1	0
1	0	1	0	1
1	1	0	0	1
1	1	1	1	1

$$S = A \oplus B \oplus C_i$$
$$= A\overline{B}\,\overline{C_i} + \overline{A}B\,\overline{C_i} + \overline{A}\,\overline{B}C_i + ABC_i$$
$$= AB + BC_i + AC_i$$

An n-bit adder can be constructed by cascading N-full adder circuits in series, connecting as shown in the figure:



4-bit ripple carry adder

In this adder the worst case delay happens when a carry generated at the least significant bit position propagates all the way to the most significant. The delay is then proportional to the number of 0 bits in the input words N and is approximated by:

$$t_{adder} = (N-1)\,t_{carry} + t_{sum}.$$

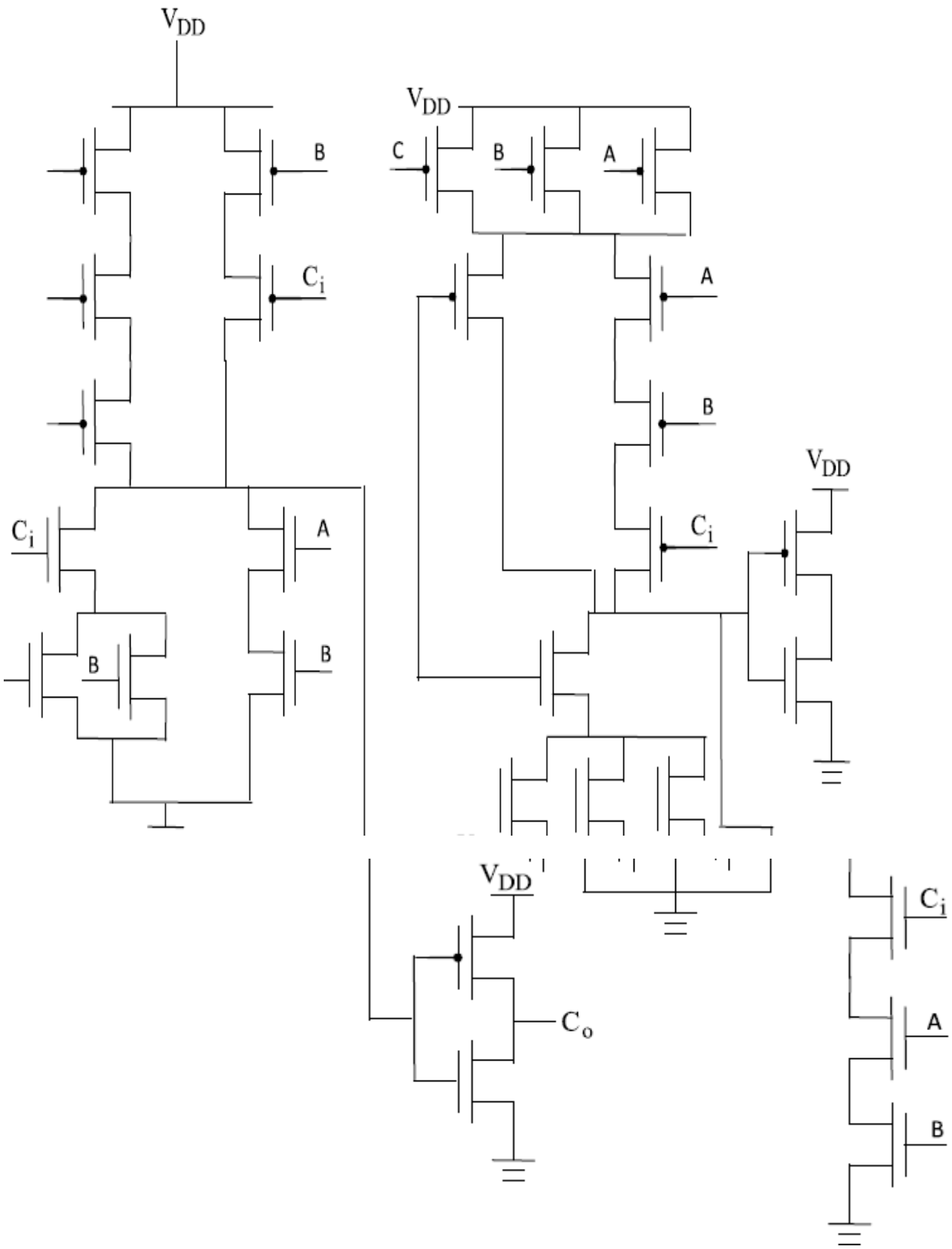
From this equation we can conclude that

1. The propagation delay of ripple carry adder is proportional to N.
2. It is far more important to optimize t_{carry} than t_{sum}.

Full Adder:-

Circuit design consideration:

The equations given above should be implemented using CMOS circuitry.



Complementary static CMOS implementation of full adder

Carry select adder and its working principle:-

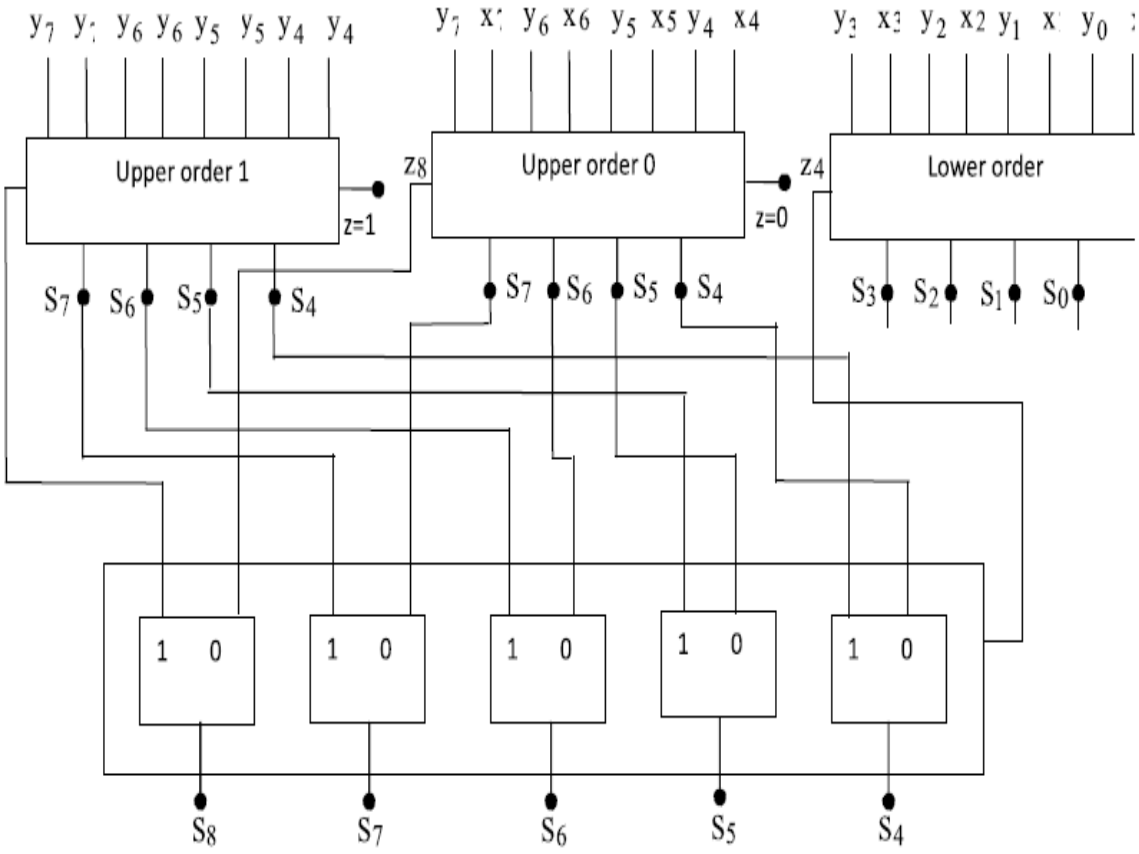
It consists of 'n' number of narrow adders (i.e. multiple) to generate fast addition. It splits the addition problem into smaller groups and special attention must be given to higher order group that adds the word cells $x_{n-1}.....x_{n/2}$ and $y_{n-1}.....y_{n/2}$. The carry bit has two possibilities $C_{n/2}=0$ or $C_{n/2}=1$. A CSA has two different adders for upper words. A multiplexer is then connected to select the proper result.

Example:-

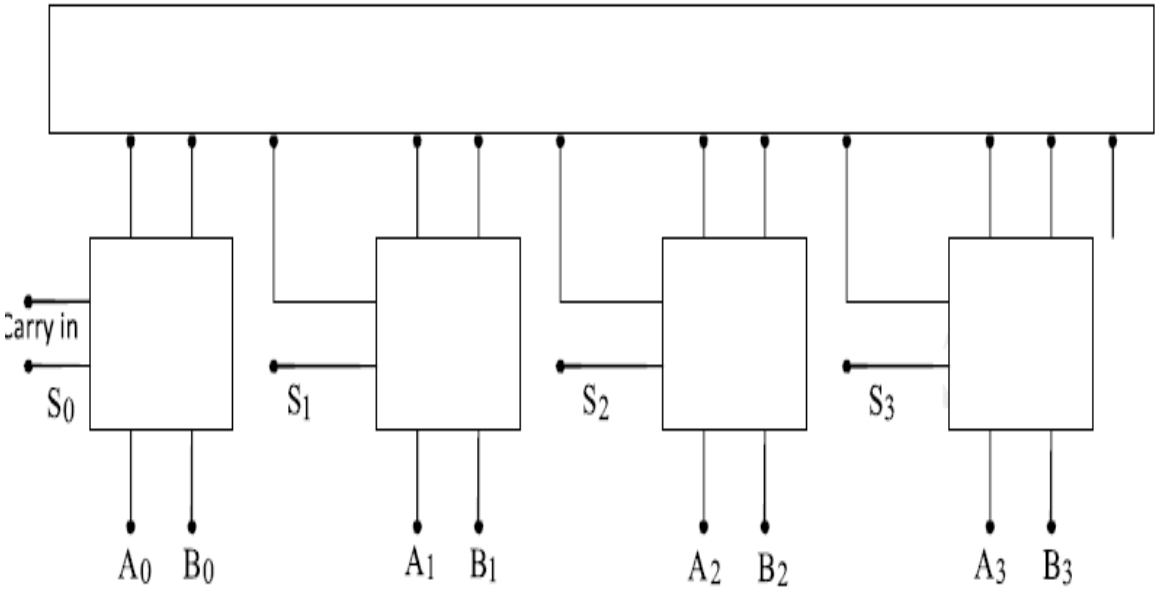
Designing of 8-bit carry select adder:-

An 8-bit CSA is splitted into two groups of each having 4-bits, the low-order bits (Lo) $x_3\ x_2\ x_1\ x_0$ and $y_3\ y_2\ y_1\ y_0$ are pared into the 4-bit adder to generate the result (sum) bits $S_3\ S_2\ S_1\ S_0$ and a carry out bit Z_4.

The higher order bits $x_7\ x_6\ x_5\ x_4$ and $y_7\ y_6\ y_5\ y_4$ are applied to two 4-bit adders. The upper order 0 adder finds the sum with a C-in of z=0 and upper order 1 finds the C-in of z-1.



Carry-Look Ahead Adder:-



The mechanism adopted by the carry look ahead adder is explained below:

In this process, here are two important variables used namely 'P' to represent "propagate" and "G" to represent "generate". A part from these variables, there are other variables used such as "c" which represents carry generated after addition process. 'A', 'B' the inputs and 'S' which represent the result of the sum of inputs. The variables 'P' and "G' process, their general equation as

$P_i = A_i + B_i$

$G_i = A_i * B_i$

The significance of P_i and G_i is as follows:

If $P_i = 1$, it means that the carry from $(i-1)^{th}$ bit is propagated or transferred to next higher bit and if $G_i = 1$, it means that a carry is generated from the i^{th} bit, after performing the sum.

The various sum and carry generated, in addition process has the general formula

$sum_i = C_i \oplus P_i \oplus G_i$

$= G_i + P_i C_i$

Expanding the above equation results in three more equations (i.e.)

$$C_{i+1} = G_i + P_i(G_{i-1} + P_{i-1}C_{i-1})$$

Opening the braces and expanding, we get,

$$C_{i+1} = G_i + P_iG_{i-1} + P_iP_{i-1}(G_{i-2} + P_{i-2}C_{i-2})$$

Multiplying P_iP_{i-1} with G_{i-2} and $P_{i-2}C_{i-2}$, we get,

$$C_{i+1} = G_i + P_iG_{i-1} + P_iP_{i-1}G_{i-2} + P_iP_{i-1}P_{i-2}C_{i-2}$$

Hence the carry at (i+1) th term is nothing but the combination of four levels of expansion. This is said to be the limit of expansion.

There are four blocks used with each block associated with 6 types of variables namely, the generate variable (G), the propagate variable (P), carry (C), sum(S) and two inputs in the form of "A" and "B" respectively. The carry generated at the first block is passed to next block, the result that is the sum in each block is computed by adding the values of inputs A and B along with previous carry. This is the chain process continue till a sole result is obtained.

ALUs:

The heart of the ALU is a four bit adder circuit. All four bit quantities are presented in parallel form and the shifter circuit has been designed to accept and shift a four bit parallel sum from the ALU.

Implementing ALU Functions:

An arithmetic and logical operations (ALU) must obviously be able to add two binary numbers (A+B), and must also be able to subtract (A-B).

From the view of logical operations, it is essential to be able to AND two binary words (A. B). It is also desirable to OR (A+B) and perhaps also detect equality.

Subtraction by an adder is an easy operation provided that the binary numbers A and B are presented in twos complement form. In this case, to find the difference A-B it is only necessary to complement B (exchange 1 for 0 and vice versa for all bits of B),add one to the number thus obtained, and then add this quantity to A using the addition process. The output of the adder will then be the required difference in twos complement form.

The standard adder equation is

Sum $S_k = \overline{H_k} C_{k-1} + H_k \overline{C_{k-1}}$

New carry $C_k = A_k B_k + H_k C_{k-1}$

Where half sum $H_k = \overline{A_k} B_k + A_k \overline{B_k}$

Consider, first the sum output if C_{k-1} is held at logical 0, then

$S_k = H_k * 1 + \overline{H_k} * 0 = H_k$

i.e. $S_k = H_k = \overline{A_k} B_k + A_k \overline{B_k}$

An Exclusive or operation

Now, hold C_{k-1} at logical 1, then

$S_k = H_k * 0 + \overline{H_k} * 1 = \overline{H_k}$

i.e. $S_k = \overline{H_k} = \overline{A_k B_k} + A_k B_k$

An Exclusive nor operation.

Next consider the, carry output of each element, first if C_{k-1} is held at logical 0, then,

$C_k = A_k B_k + H_k * 0 = A_k B_k$

An AND operation

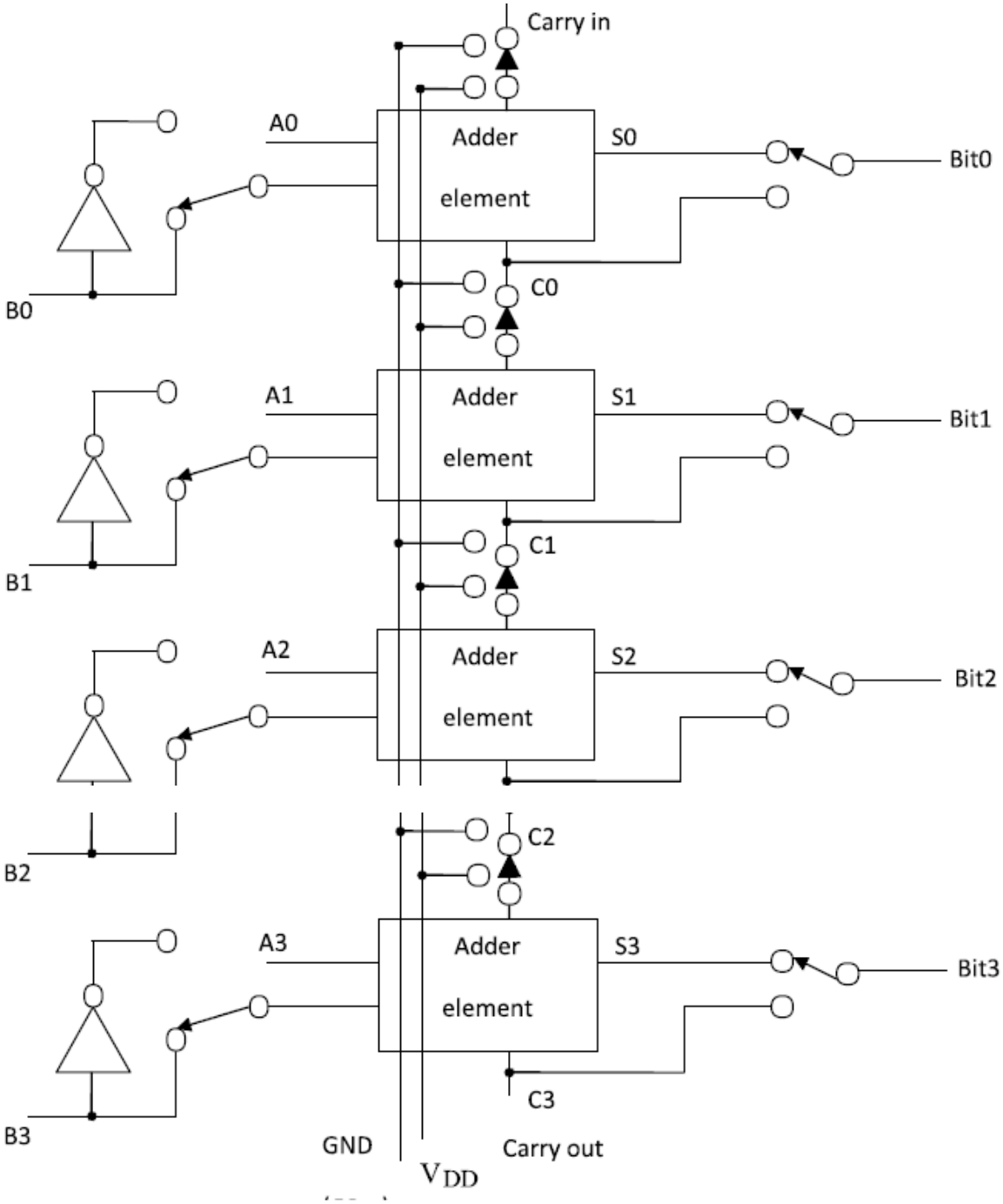
Now if C_{k-1} is held at logical 1, then,

$C_k = A_k B_k + H_k * 1 = A_k B_k + \overline{A_k} B_k + A_k \overline{B_k}$

Therefore $C_k = A_k + B_k$

An OR operation

Thus it may be seen that suitable switching of the carry line will give the ALU logical functions. This can be observed in below figure.



4-bit ALU

MULTIPLIERS:

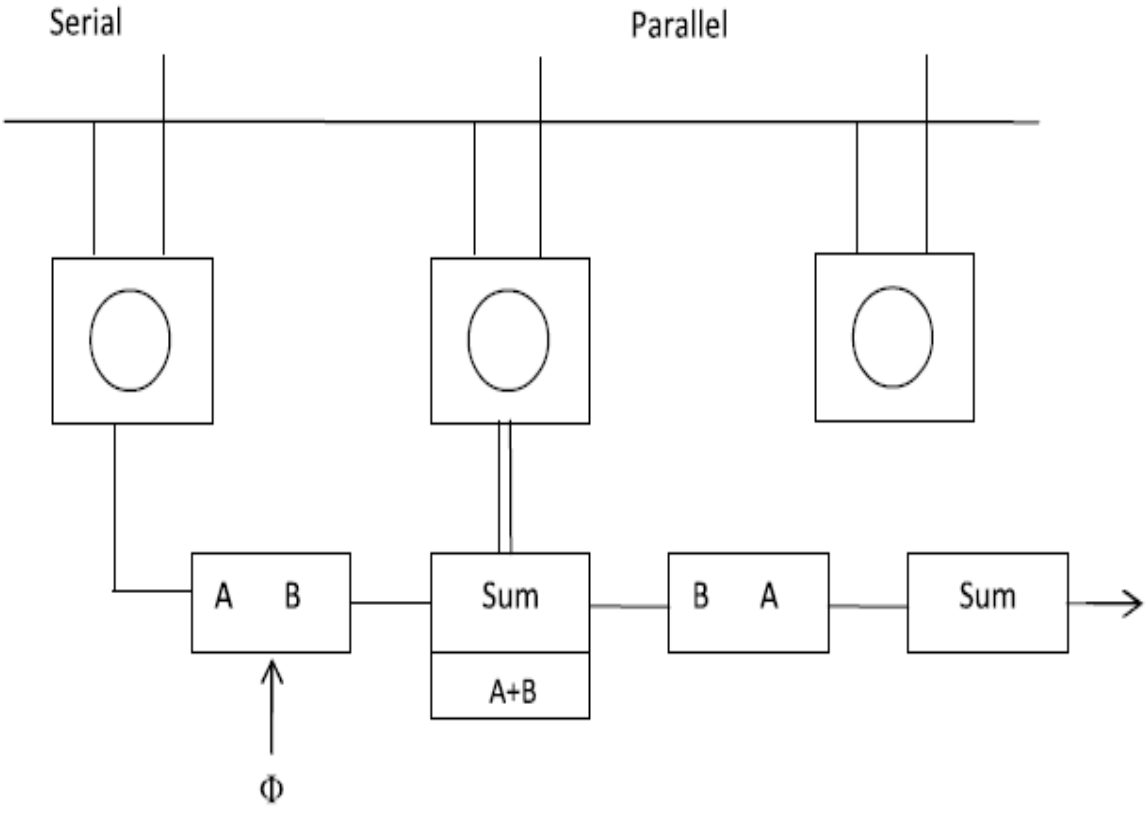
The multiplication is performed by multiplying one digit of the multiplier by the multiplicand at each and every step and then adding the result by shifting the number of bits to the product (partial)

Example:-

Multiplicand		0	1	0
Multiplier		0	1	1
Partial product	*			
	+	0	1	0
	0	1	0	
	0	1	1	0
		0	0	0
0	0	1	1	0

Serial Parallel Multiplier:-

The multiplication of binary two bits can be performed by the AND function. The basic device used for multiplication is serial-parallel multiplier. Here, the n-bit multiplier in serially where as, the y-bit multiplicand is passed parallely.



Initially the multiplier is connected to the LSB (least significant bit) and it is also connected by at least y zeroes. At the end of the multiplier chain, the output appears serially. The, sum box includes a full adder and a register to save the carry. The continuous summation of chain units and registers has the facility of the shift and add operation. One of the best algorithm for signed multiplication is Baught-Wooley multiplier. It helps to provide all the partial products to achieve positive sign bits.

The formula of the multiplier A is

$$A = a_{n-1}2^{n-1} + \sum_{j=0}^{n-2} a_j 2^j$$

Where n=number of bits.

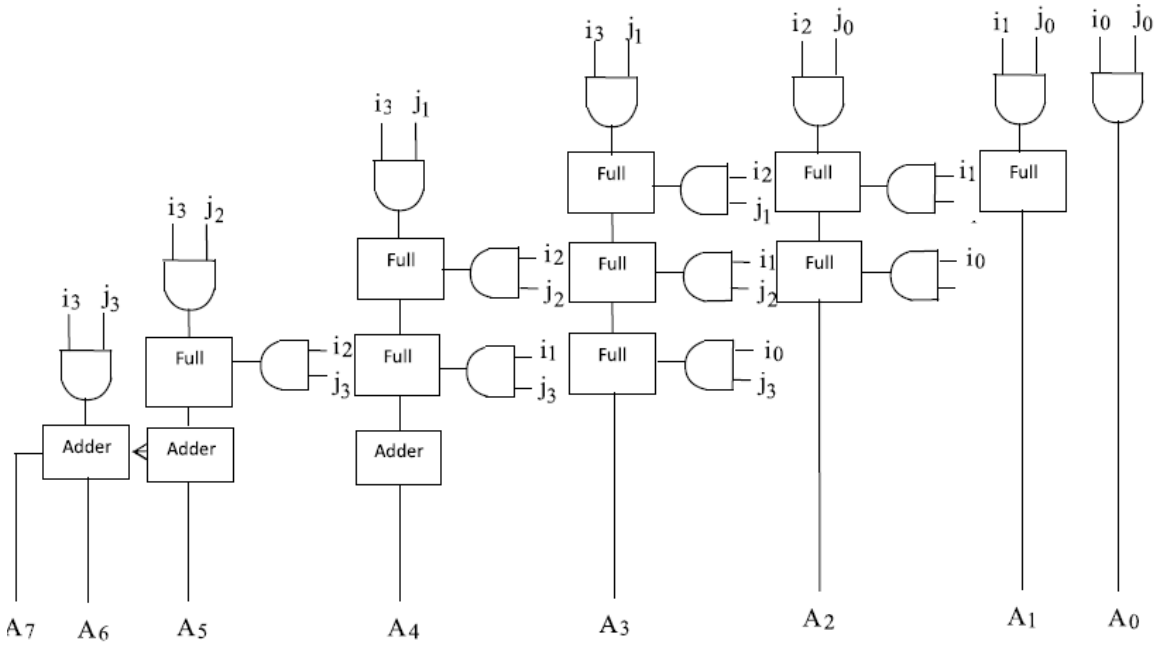
The formula of multiplicand B is

$$A = B_{2n-1}2^{2n-1} + \sum_{j=0}^{2n-2} B_j 2^j$$

The above formula can also be expanded to generate the partial products. Each partial product is generated with AND functions and later all these products are added.

Unsigned Array Multiplier:

The structure of unsigned array multiplier:

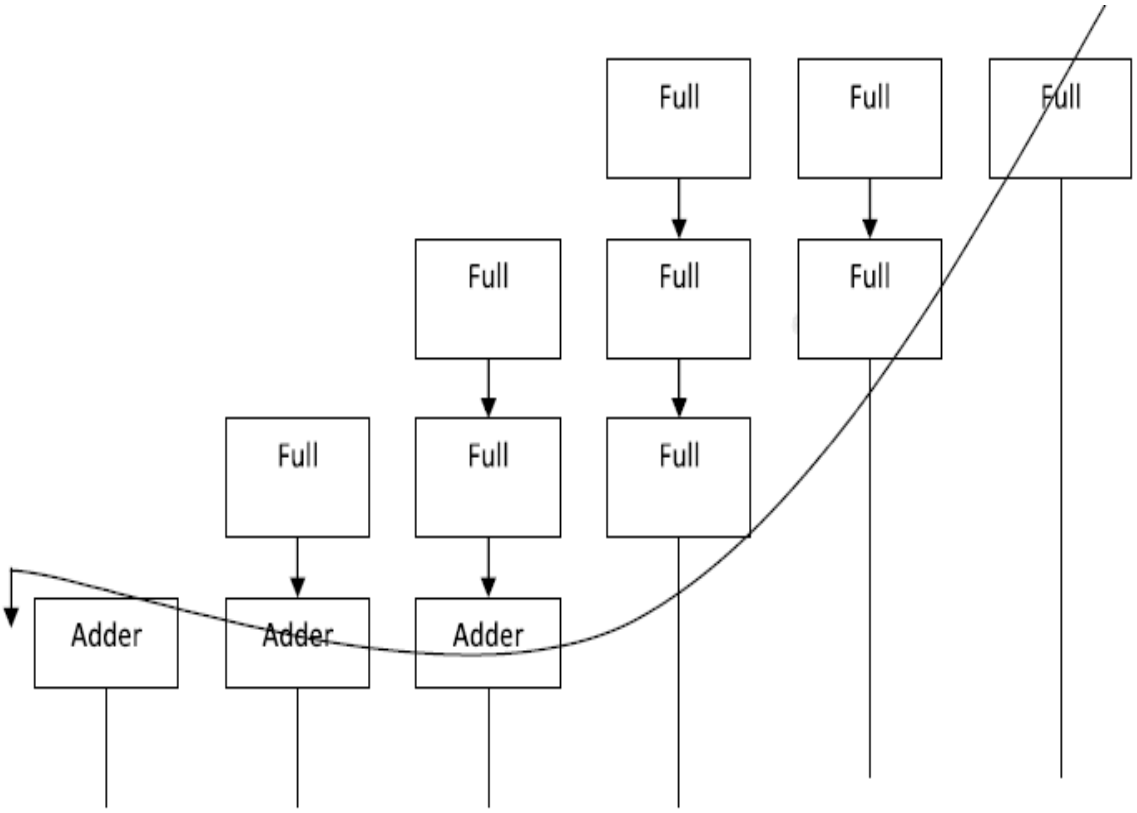


In normal multiplication, we follow a format in which, a series of partial products are obtained. In regular steps, these partial products are moved left at one bit position, causing 1 bit free space at right. After the multiplication process is completed, the partial products are added up, to obtain the derived result. The result of each iteration is released down and a single bit value is provided to the adjacent full adder box. The full adder box takes up the i_0

Value from the previous adder and computes the multiplication process using j_1 and also supplies the value to the next full adder. Later the full adder box takes in the $i_o j_1, i_1 j_o$ values coming from their respective A and D gates, giving rise to partial product. The partial product obtained is also supplied to the adjacent adder.

The process is continued till the desired product is obtained. The last adder in the circuit possess a carry chain where as the initial iterations are performed by full adders. The full adders intakes three1-bit input and produces two 1-bit outputs.

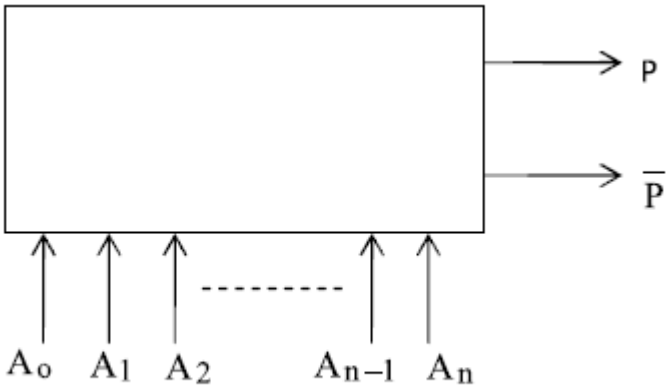
When the critical path of the circuit is drawn, a trajectory path is reflected as shown in the figure below



PARITY GENERATOR:

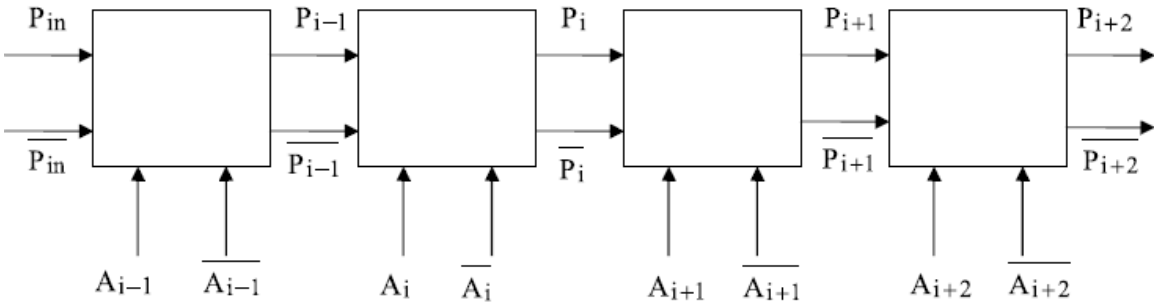
A circuit is to be designed to indicate the parity of a binary number or word. The requirement is indicated in figure 1 below for an (n+1)-bit input.

Since the number of bits is undefined, we must find a general solution on a cascadable bit wise basis, so that, 'n' can have any value. A suitably regular structure is set out in figure 2



Note: P = 1 Even number of 1's at input

Parity generator basic block diagram



Parity Generator-Structured Design Approach

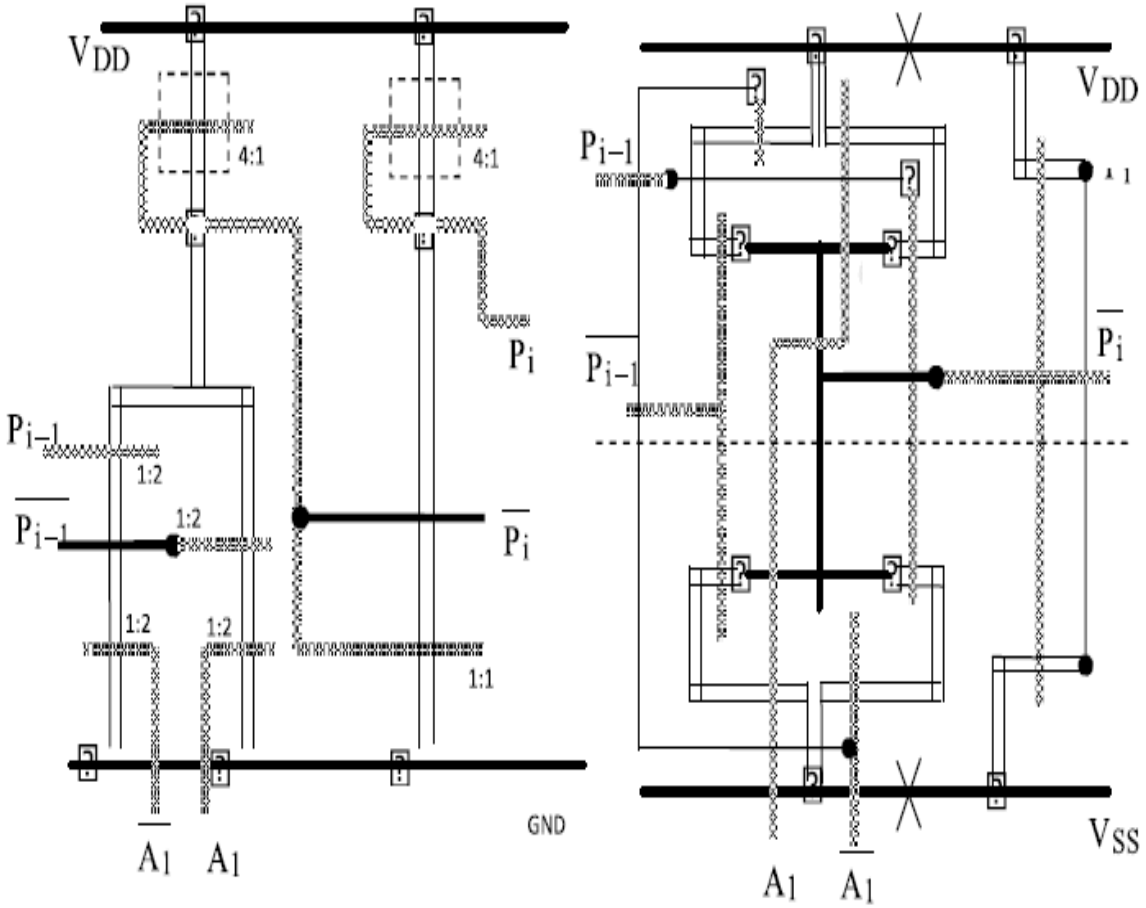
A little reflection will readily say the requirements are:-

Ai = 1 parity is changed, $P_i = \overline{P_{i-1}}$

Ai = 0 parity is unchanged, $P_i = \overline{P_{i-1}}$

A suitable arrangement for such a cell is given in this stick diagram in the figure (3) (nMOS) and figure(4) (cmos)

The cell boundary may be chosen in each case so that cells may be cascaded at will.



COMPARATORS:

The comparison of two numbers is an operation that determines if one number is greater than, less than or equal to the other number. A magnitude comparator combinational circuit that compares the two numbers, A and B, and determines their relative magnitudes. The outcome of the comparison is specified by three binary variables that indicates whether A > B, A = B or A < B.

The algorithm is explained next.

Consider 2 numbers A and B with 4 digits each. The coefficients of the numbers are written as below:

$$A = A_3 A_2 A_1 A_0$$
$$B = B_3 B_2 B_1 B_0$$

Where each subscribed letter represents one of the digits in the number. The 2 numbers are equal if all the pairs of significant digits are equal, i.e., if $A_3 = B_3$ and $A_2 = B_2$ and $A_1 = B_1$ and $A_0 = B_0$. When the numbers are binary, the digits are either one or zero and the equality relation of each pair of bits can be expressed logically with an equivalence function:

$$x_i = A_i B_i + A_i' B_i'$$

Where $x_i = 1$ only if the pair of bits in position 'i' are equal, i.e. , if both are 1's or both are 0's.

The equality of the 2 numbers A and B is displayed in a combinational circuit by an output binary variable that we designate by the symbol (A = B).

This binary variable is equal to one if the input numbers A and B are equal; and it is equal to zero otherwise. For the equality condition to exist, all x_i variables must be equal to 1. This dictates an AND operation of all variables.

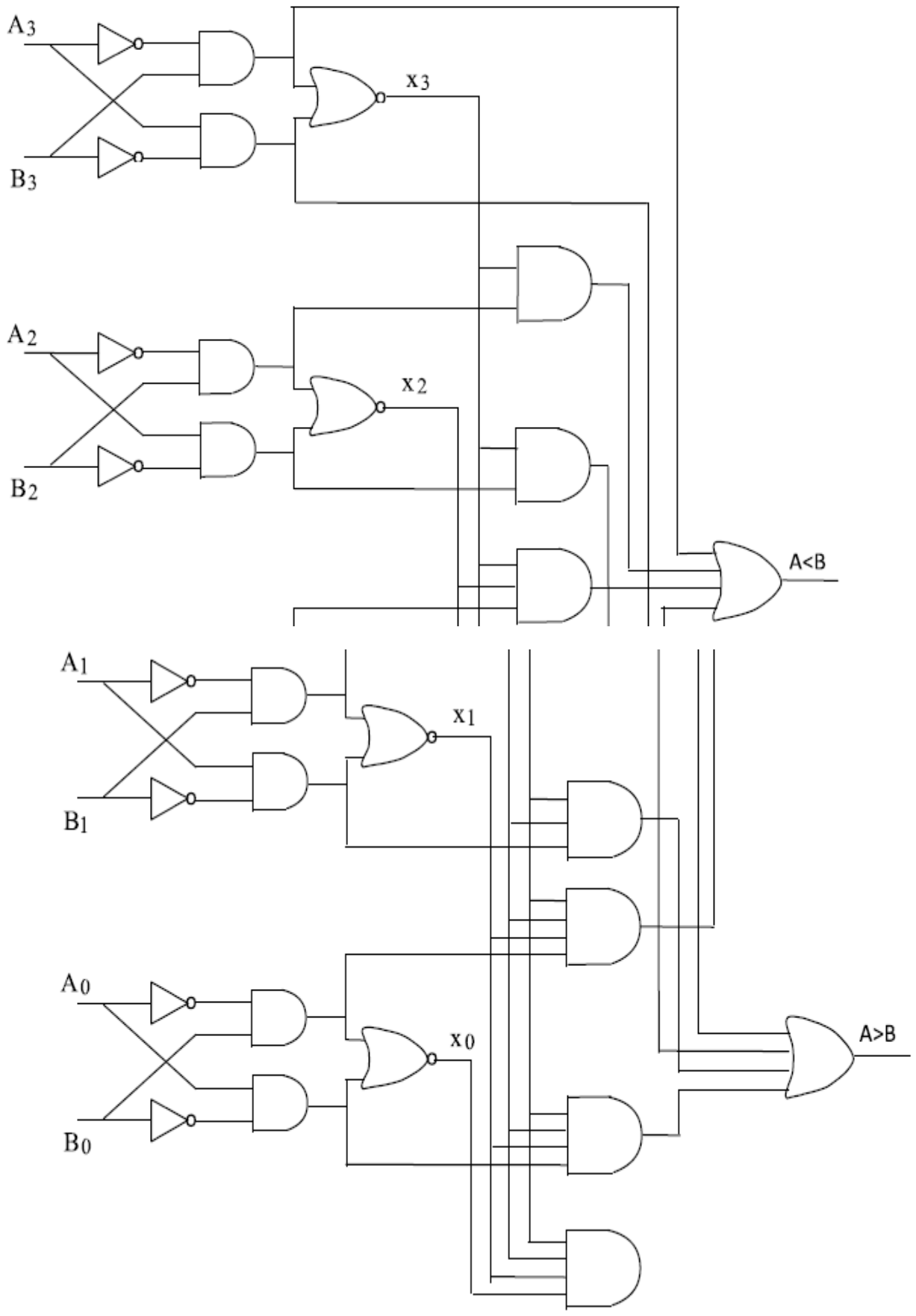
$$(A = B) = x_3 x_2 x_1 x_0$$

To determine if A is greater than or less than B, we inspect the relative magnitudes of pairs of significant digits starting from the most significant position. The sequential comparison can be expressed logically by the following 2 Boolean functions:

$$(A > B) = A_3 B_3' + x_3 A_2 B_2' + x_3 x_2 A_1 B_1' + x_3 x_2 x_1 A_0 B_0'$$

$$(A < B) = A_3' B_3 + x_3 A_2' B_2 + x_3 x_2 A_1' B_1 + x_3 x_2 x_1 A_0' B_0$$

The "unequal" outputs can use the same gates that are needed to generate the equal output. The logic diagram of the 4-bit magnitude comparator is shown in the below figure:

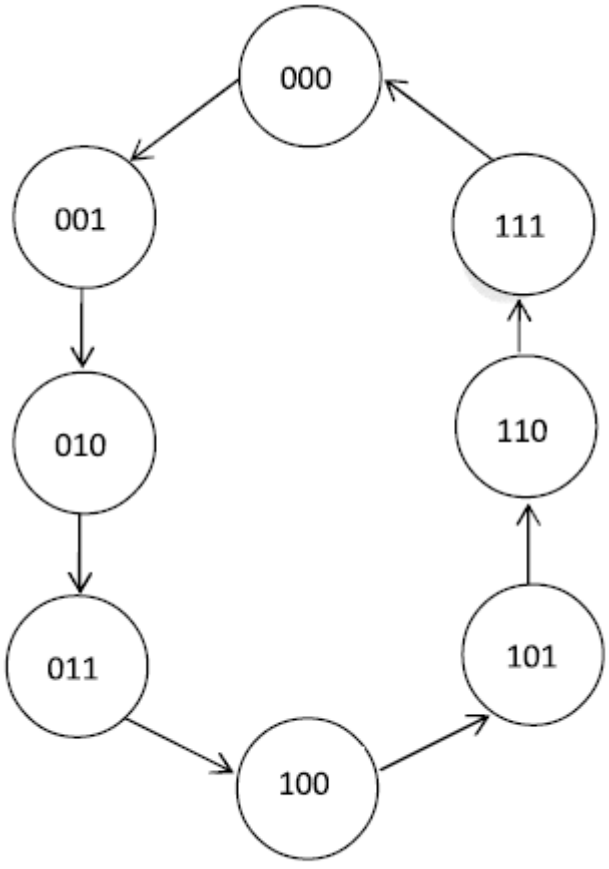

A3
B3
x3
A2
B2
x2
A<B
A1
B1
x1
A0
B0
x0
A>B

COUNTERS:

A sequential circuit that goes through a prescribed sequence of states upon the application of input pulses is called a counter. The input pulses, called count pulses, may be clock pulses ar they may originate from an external source. In a counter the sequence of states may follow a binary count or any other sequence of states.

A counter that follows the binary sequence is called a binary counter. An n-bit binary counter consists of 'n' flip-flops and can count in binary from 0 to 2^{n-1}. The state diagram of a 3-bit counter is shown in figure below:

The flip-flop outputs repeat the binary count sequence with a return to 000 after 111. the outputs are directly specified by the present states of the flip-flops. The next stste of a counter depends entirely on its present state, and the state transition occurs every time the pulse occurs.



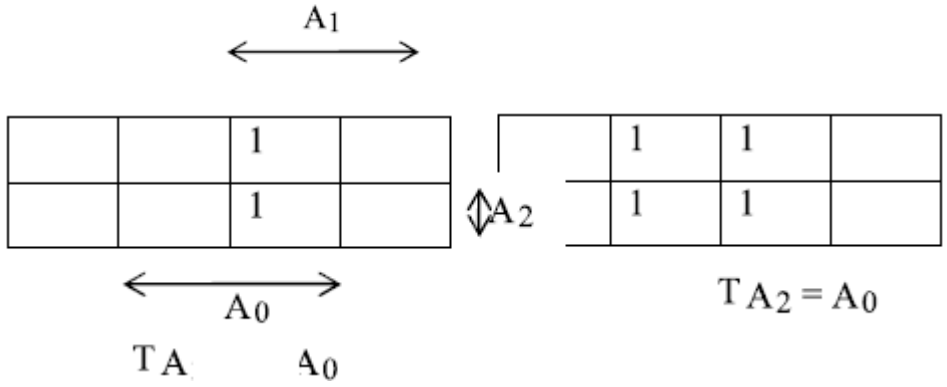
State Diagram of 3-bit Binary Counter

The excitation table for 3-bit binary counter is:

The three flip-flops are given variable designations A_2, A_1 and A_0. Binary counters are most efficiently constructed with T flip-flops (or JK flip-flops with J and K tied together). The flip flop excitation for the T inputs is derived from the excitation table for the T flip flop and from inspection of the state transition from present to next state. Consider the flip-flop input entries for row 001. The present state here is 001 and the next state is 010, which is the next count in the sequence. Comparing these two counts, we note that A_2 goes from 0 to 0;so TA_2 is marked with a 0 because flip flop A_2 must remain unchanged when a clock pulse occurs. A_1 goes from 0 to 1; so TA_1 is marked with a 1 because this flip flop must be complemented in the next clock pulse. Similarly A_0 goes from 1 to 0, indicating it must be complemented, so T A_0 is marked with 1.

Present State		**Next State**		**Flip-Flop Inputs**	
A_2	A_1	A_2	A_1	A_2	A_1
A_0		A_0		A_0	
0	0	0	0	0	0
0		1		1	
0	0	0	1	0	1
1		0		1	
0	1	0	1	0	0
0		1		1	
0	1	1	0	1	1
1		0		1	
1	0	1	0	0	0
0		1		1	
1	0	1	1	0	1
1		0		1	
1	1	1	1	0	1
0		1		1	
1	1	0	0	1	1
1		0		1	

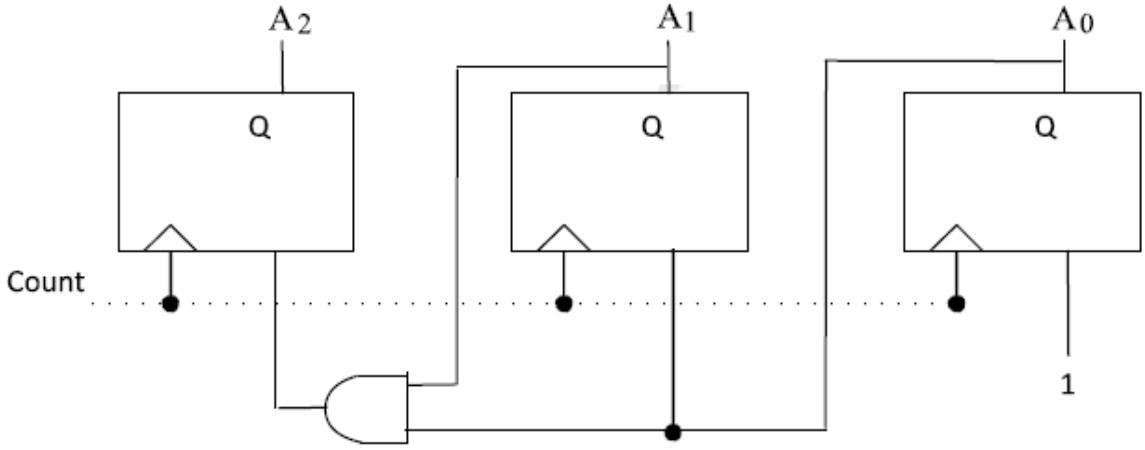
The flip-flop input functions from the excitation tables are simplified in the maps of the following figure.



1	1	1	1
1	1	1	1

$$T_{A_0} = 1$$

The Boolean functions listed under each map specify the combinational circuit part of the counter as shown in figure below:



Logic Diagram of a 3-bit Binary Counter

Counter with non-binary sequence:-

A counter with n flip-flops may have a binary sequence of less than 2^n states. For example consider the counter mentioned in the table below:

Present state			Next state			Flip-flop Inputs					
A	B	C	A	B	C	J A	K A	J B	K B	J C	K C
0	0	0	0	0	1	0	X	0	X	1	X
0	0	1	0	1	0	0	X	1	X	X	1
0	1	0	1	0	0	1	X	X	1	0	X
1	0	0	1	0	1	X	0	0	X	1	X
1	0	1	1	1	0	X	0	1	X	X	1
1	1	0	0	0	0	X	1	X	1	0	X

Inputs KB and KC have only 1's and X's in their columns, so these inputs are always equal to 1.

The count has a repeated sequence of six states, with flip flops B and C repeating the binary count 00, 01, 10 while flip flop A alternates between 0 and 1 every three counts. The count sequence is not straight binary and two states, 011 and 111, are not included in the count.

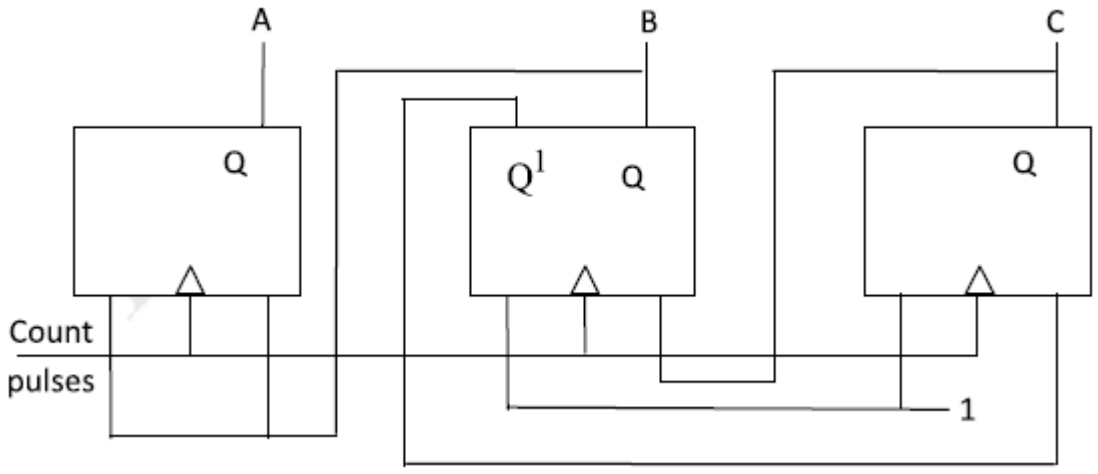
The flip flop inputs other than KB and KC can be simplified using min terms 3 and 7 as don't care conditions the simplified functions are

$JA = B$ $KA = B$

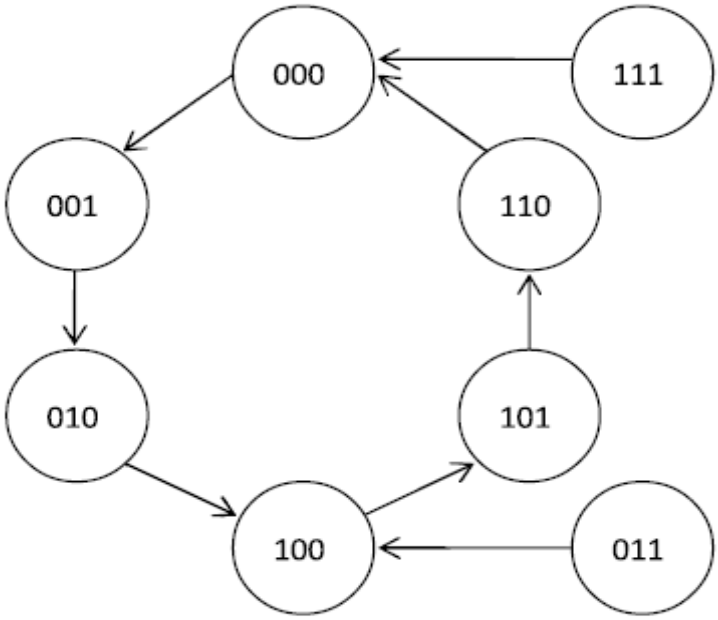
$JB = C$ $KB = 1$

$JC = B^I$ $KC = 1$

The logic diagram of the counter is shown below:



State diagram of the counter is shown below:



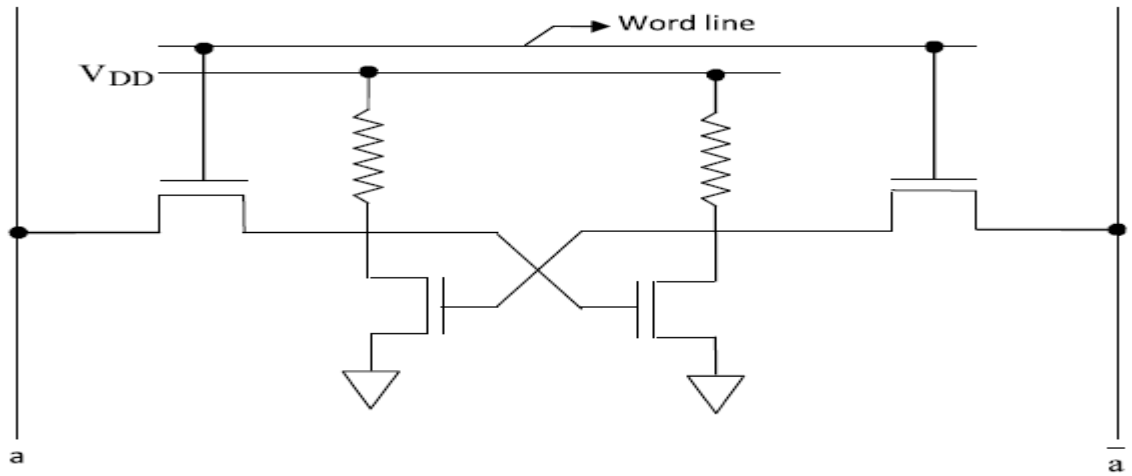
If the circuit happens to be in the state 011 because of an error signal, the circuit goes to the state 100, after the application of a clock pulse.

The state diagram including the unused states is shown in the above figure. If the circuit ever goes to one of the unused states because of an error, the next count pulse transfers it to one of the valid states and the circuit continues to count correctly.

HIGH DENSITY MEMORY ELEMENTS:-

SRAM:-

The circuit diagram of the resistive load static RAM cell is given below:



The two lines are named a and $\bar{a}$ on the either sides of the circuit. There are another two lines over the top of the circuit representing V_{DD} and word line respectively.

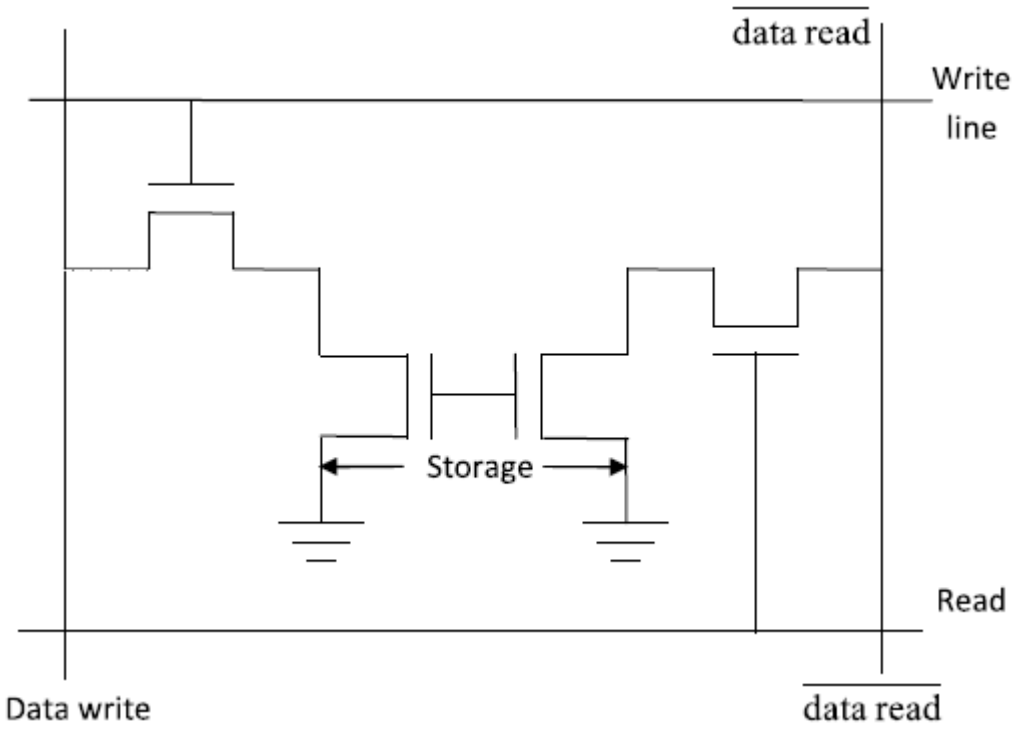
The working of the above cell is divided into three phases. Phase-I deals with the normal case (i.e.) whenever the word line is equal to '0', this means the cell is isolated or not working.

The phase-II deals with write operation. The write operation is performed by providing necessary voltages to the lines a and $\bar{a}$ respectively. The word line is given a value 1, this causes storage of desired value. The lines a and $\bar{a}$ posses the capacitance, which is greater then than the transistors, also it is sufficient in moving the transistors to flip state.

Phase-III deals with read operation. During this operation the two lines a and $\bar{a}$ are charged to V_{DD}. Here two conditions occurs (i.e.) if the right transistors output value is '0', causes the '$\bar{a}$' line to be drained, at the same time the line 'a' will remain high.

Similarly, if the left side transistors output is '0' then 'a' line gets drained, at the same time the '$\bar{a}$', line remains high. In this way the stored value is read.

Four transistor DRAM cell with storage nodes:-



As seen from the figure above there are four dynamic RAM cells in which two DRAM's are placed on either ends of the circuit whereas other two DRAM's occupies the central position of the circuit. Also there are two additional lines provided on the top as well as bottom of the circuit.

Dynamic RAM can be used for reading as well as writing. Hence working of the cell is nothing but either retrieving the data from the cell or storing the data into the cell (known as writing).

Writing Data To The Cell:-

During writing process, the data write line consists of the data to be written and write line is set to '1', this automatically makes the line read to zero. Finally, the charge present at the node +1 and the data write forces +1 to the desired value, hence the data gets stored.

Reading Data From The Cell:-

Reading process is simple. During this process, data read line is activated by providing it with a value '1', similar case is also observed with the read line (i.e. it is also activated by providing a value '1'). This automatically deactivates the write line (i.e. '0' is provided to it). Finally, the storage node +1 pull-down the read data line causing the stored data to be flown out through the storage node.

UNIT-V

Contents:

- PLAs
- FPGAs
- CPLDs
- Standard Cells
- Programmable Array Logic
- Design Approach
- Parameters Influencing Low power Design.

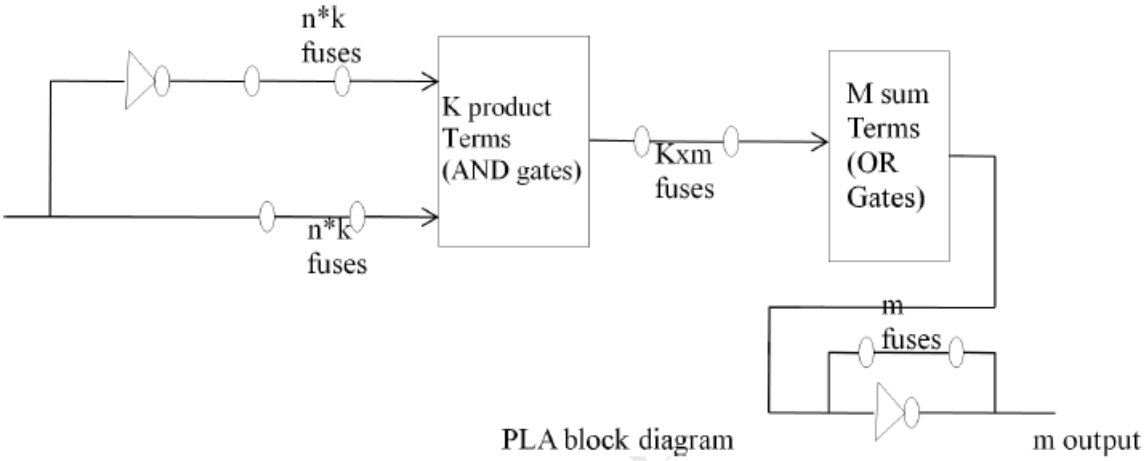
CMOS TESTING:

- CMOS Testing
- Need for testing
- Test Principles
- Design Strategies for test
- Chip level Test Techniques
- System-level Test Techniques
- Layout Design for improved Testability.

UNIT-5

PLA's:

PLA is abbreviated form of programmable logic array.IN PLA we have a group of 'and' gates in the first stage, each of wich can be programmed to generate a product term of the input variables the AND and OR gates inside the PLA are initially fabricated with fuses among them the specific Boolean functions are implemented in sum of products form by blowing appropriate fuses and learning the desired connections A block diagram of PLA is shown in the figure below:

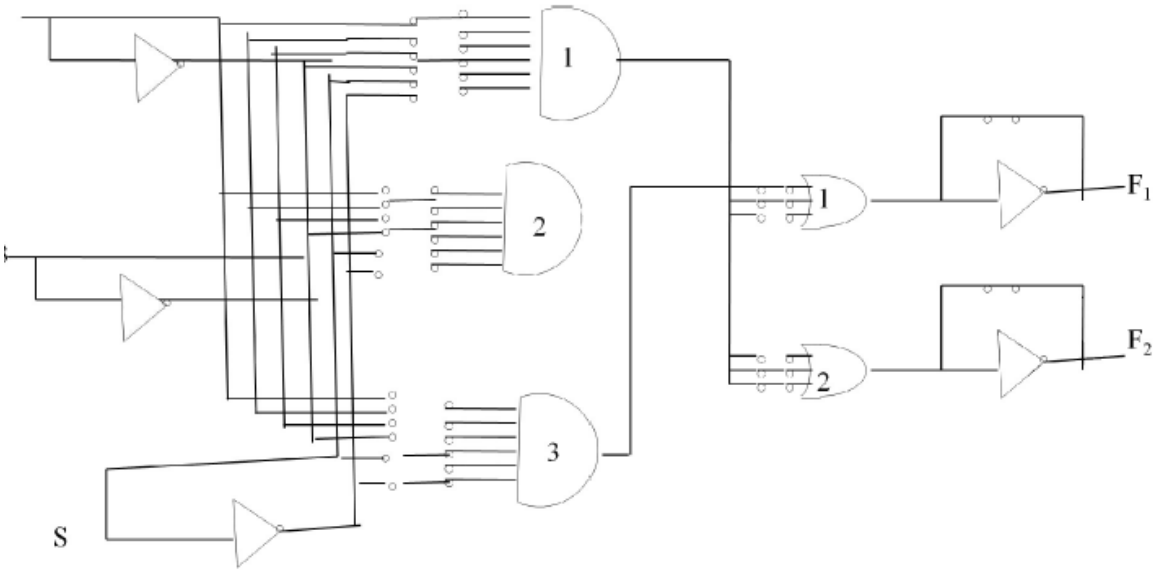


PLA block diagram m output

The diagram consists of n input m, inputs, k inputs terms, and m sum terms. The product terms constitute a group of m OR gates. Fuses are inserted between all n inputs and their complement values to each of the AND gate. Fuses are also provided between the outputs of the AND gates and the inputs of the OR gate. Another set of fuses in the output inverters allows the output function to be generated either in the AND-OR form or in the AND-OR-INWERT form.

The size of the PLA is specified by the numbers of inputs, the numbers of product terms, the number of product terms, and the number of outputs (the numbers of sum terms is equal to the numbers of outputs) A typical PLA has 16 inputs, 48 product terms, and 8 outputs. The number of programmed fuses is 2^n *m.
The internal construction of a PLA having three inputs is shown below. It has three product terms and two outputs.

PLA with three inputs, three product terms, and two outputs; it implements The combinational circuit specified in Fig.3 below



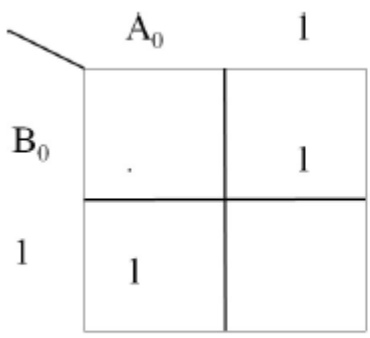
As with a ROM, the PLA may e mark – Programmable or field Programmable. With Programmable PLA, the customer must submit a PLA Program table to the manufacturer. This tale is used y the vendor to produce a custom – made PLA that has the required internal paths between inputs and outputs. A second type programmable logic array in which the program tale is not submitted by the customer.

We have another type in PLA such as Bipolar PLA and n MOS PLA. In Bipolar PLA the Switches are replaced by diodes and in n MOS PLA the replaced y diodes and in n MOS PLA the switches are n MOS transistors.

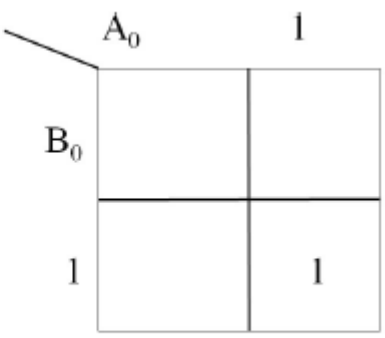
Half Adder Circuit Implementation Using PLA:-

Truth table of a half adder circuit is given below:-

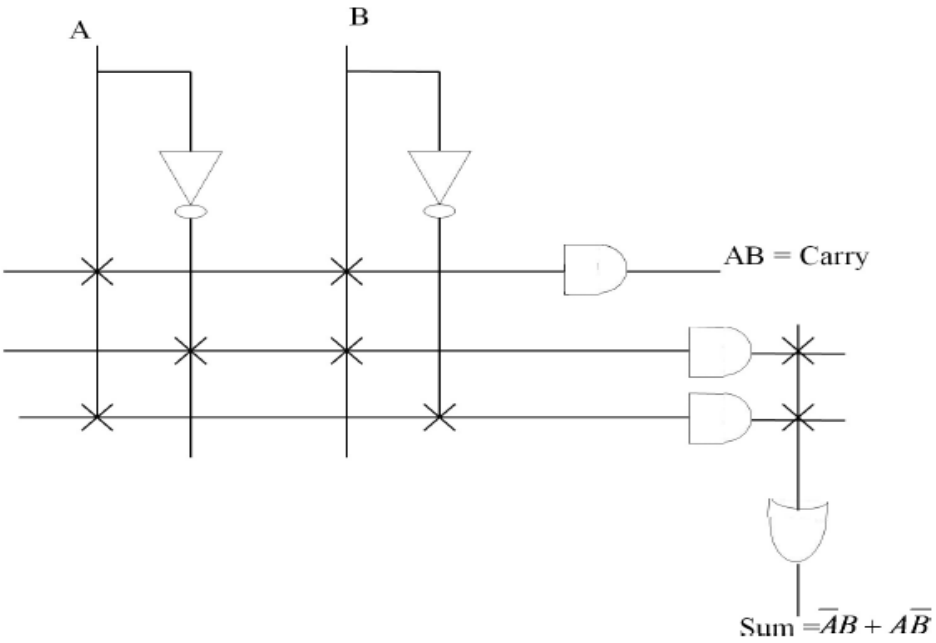
A	B	Sum	Carry
0	0	0	0
0	1	1	0
1	0	1	0
1	1	1	0



Sum = $\overline{A}B + A\overline{B}$



Carry = AB

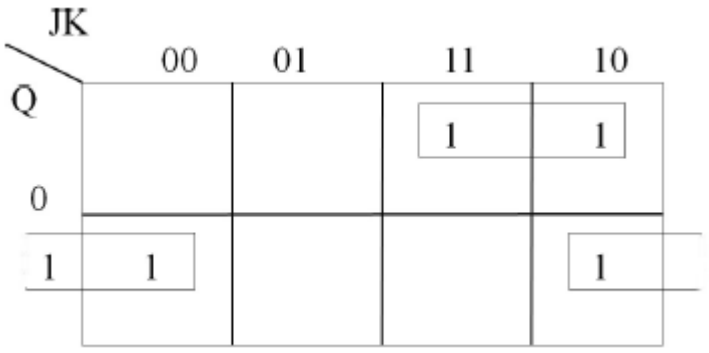


Implementation Of JK Flip Flop Using PLA:-

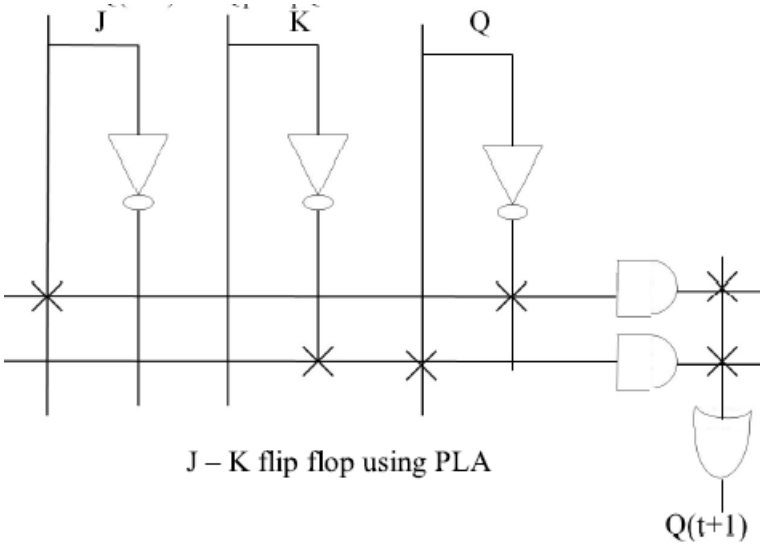
The truth tale of JK slip flop is given below:-

Q(t)	J	K	Q(T+1)
0	0	0	0
0	0	1	0
0	1	0	1
0	1	1	1
1	0	0	1
1	0	1	0
1	1	0	1
1	1	1	0

The corresponding characteristic equation is obtained by



$Q(t+1)=JQ^1+K^1Q$



J – K flip flop using PLA

FPGAS:

- A field programmable gate array is a block of programmable logic that can implement multi – Level Logic functions.
- F P G A are most commonly used as separate commodity chips that can e programmed to implement larger functions that can e fit into a programmable logic device.
- An FPGA lock must implement both combinational logic functions and interconnect to be able to construct Multi – Level Logic functions.
- All FPGAS has certain key elements in common such as a regular array of basic logic cells that are configured using programming technology.
- The chip inputs and outputs use special I/O logic cells that are different from the aspic Logic cells.
- A Programmable inter connect scheme forms the wiring between the two types of Logic cells. Finally the designer uses custom software, tailored to each programming technology and FPGA architecture, to design and implement the Programmable inter connections.

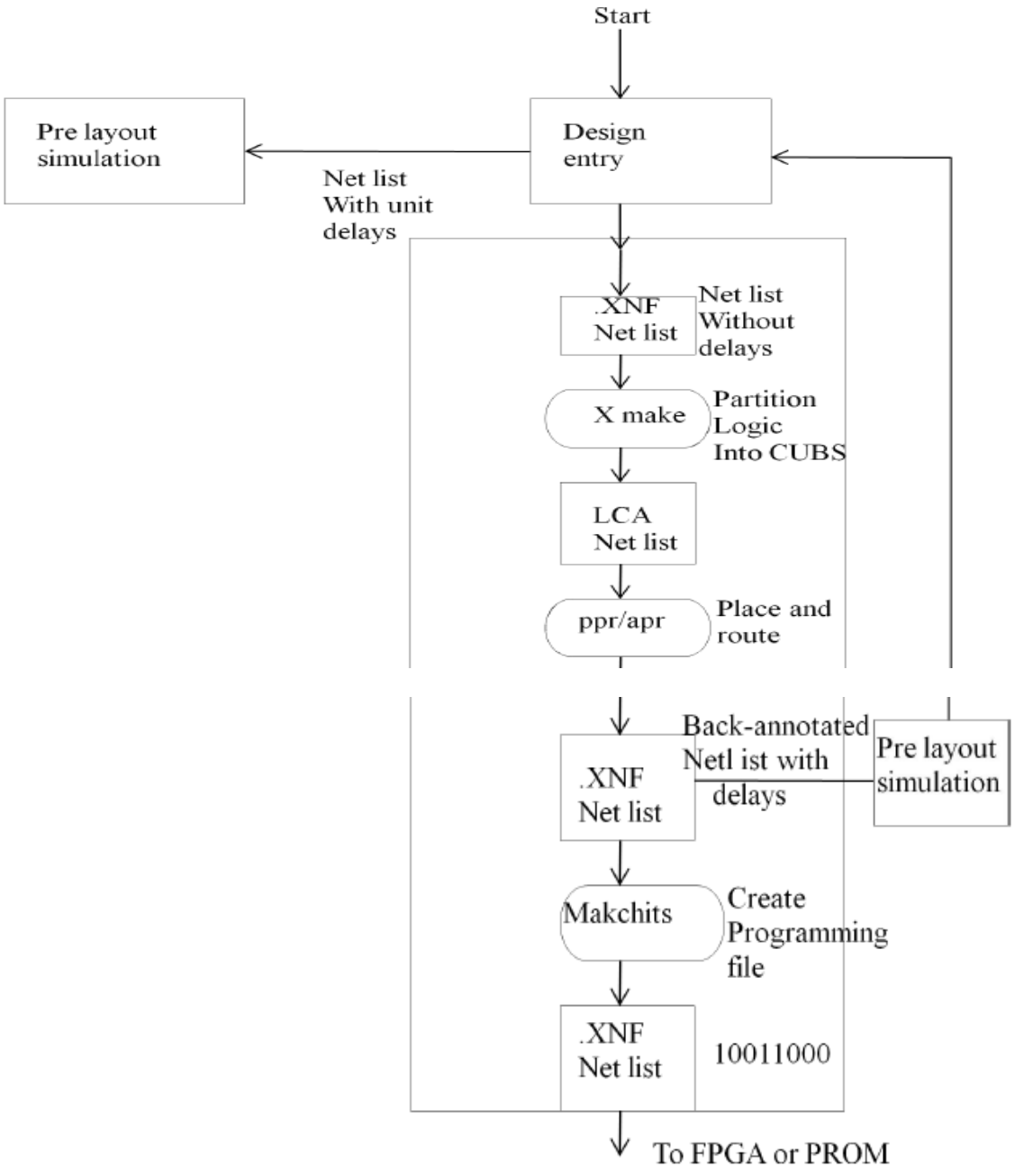
FPGA Design:-

The FPGA design is similar to the sequence for designing any ASIC or IC. The first step consists of design enky and generating a net list, the next step is simulation. Two types of simulator are normally used for FPGA design:-

(1) It is a logic simulator for behavioral, functional and timing design. This tool can catch any design errors. The designer provides input waveforms to the simulator and checks to see the outputs expected Logic path delays are only estimates. Wiring delays are known after Physical design. Then the designers add post layout timing information to the post Layout net list

(2) The second type of simulator is a timing analysis tool. A timing analyser is a Static Simulator and removes the need for input wave forms. It checks the critical paths that Limit the speed of operation.

The Typical Xilinx FPGA design flow is shown in figure below:-



FPGA design

FPGA are an extension of Mark Programmable gate arrays (MPGAS) commonly referred to as Gate arrays. Gate arrays consist of cellular rows of n MOS and PMOS transistors with provisions for interconnecting the transistors within the

cells to form gates and for routing signals between the cells. FPGAS provide logic locks for implementation of the required functions.

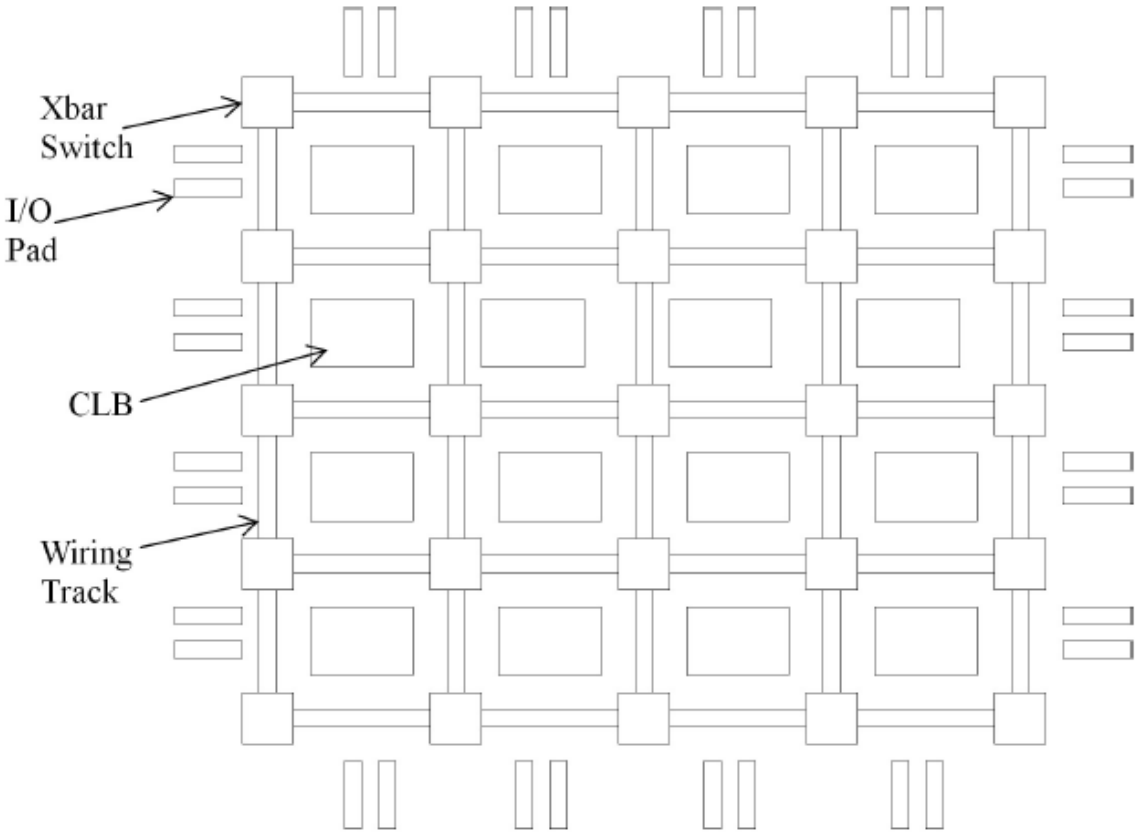


Fig (3): 2-D Array of Cells for an FPGA

The general structure of FPGA is illustrated in figure above. The figure depicts the essential elements of an FPGA organized as a $2-D$ array of cells. An FPGA consists of the following main types of resources.

1. A rectangular array of configurable Logic locks (CLBS) capable of implementing a variety of Logic functions.
2. Wiring tracks to route signals between the cells.
3. X bar switches to connect horizontal and vertical wires and
4. Input / Output pads for signal conditioning at the chip input and output pins.

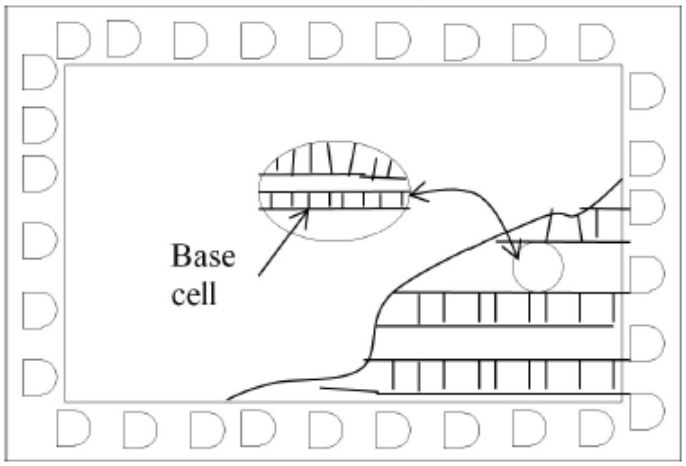
Applications of FPGAS:

- FPGAS can e applied to a very wide range of applications including random Logic, integrating multiple SPLDS, Device controllers, communication encoding and filtering small to medium sized with SRAM blocks and many more
- Prototyping of designs later to be implemented in gate arrays.
- Emulation of entire hardware systems. Emulation would entail many FPGAS connected by inter connect

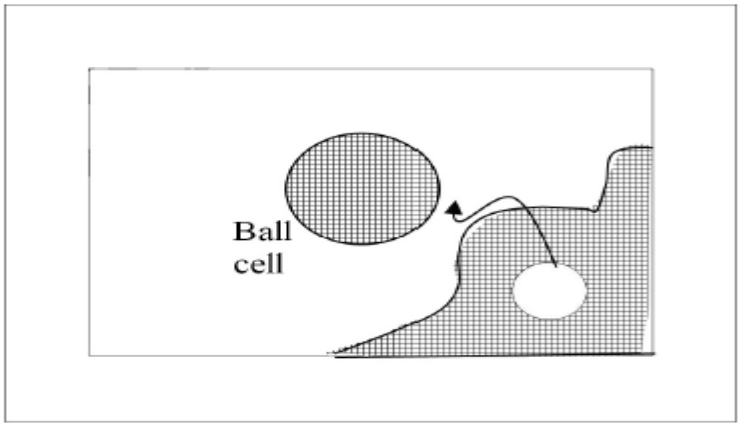
- FPGAs are also used in custom computing machines. This involves using the programmable parks to "execute" software, rather than compiling the software for execution on a regular CPU.

GATE ARRAYS BASED ASICS:

(a) Channeled gate array:- A channeled gate array uses row of cells separated y channels used for the inter connection. The figure shown below shows the channeled gate array.



Channeled Gate Array

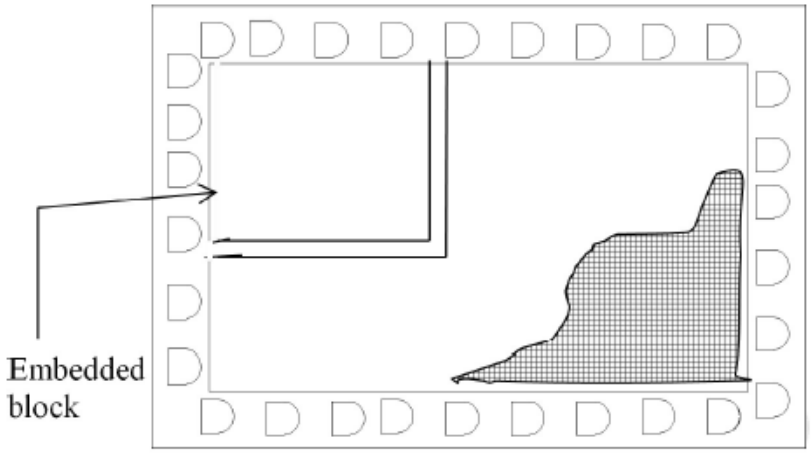


Channel Less gate array

The major features of Channel Lees gate arrays are

1. Only some (the top few) marks layers are customized the interconnect
2. The time for manufacturing lead is between two days to two weeks.

(c) Structured gate array:- It is also known as embedded gate array. One of the back draw of the masked gate array (MGA) is the fixed gate array base cell, which makes memory implementation difficult and inefficient



Structured Gate Array

In the structured gate array, some of the IC area is set aside and dedicated to a specific function. This embedded area can either contain a different base cell which is suitable more for memory cell building or it can contain a complete circuit lock such as micro controller.

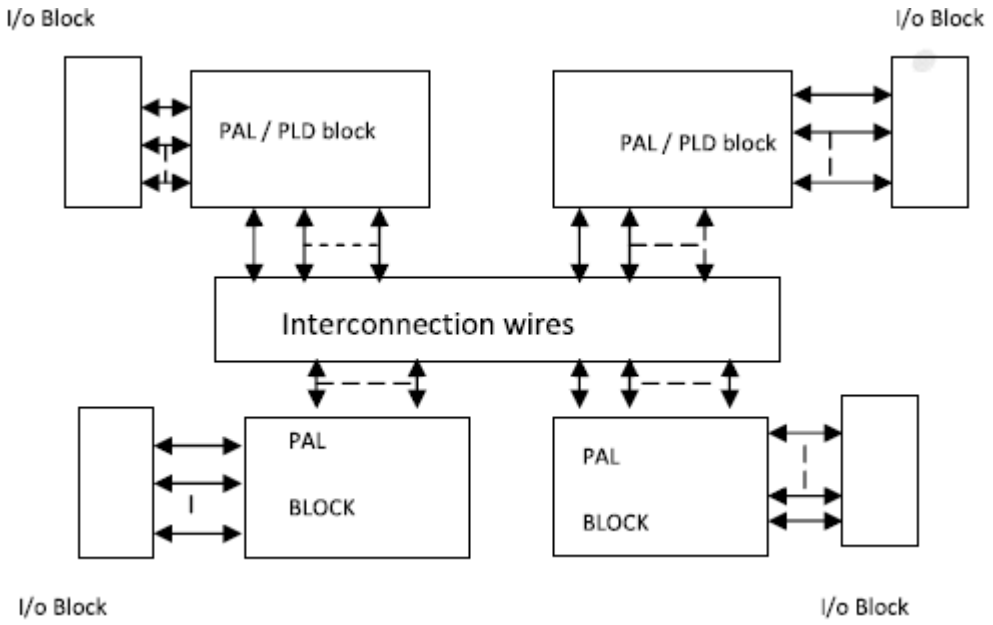
CPLD:

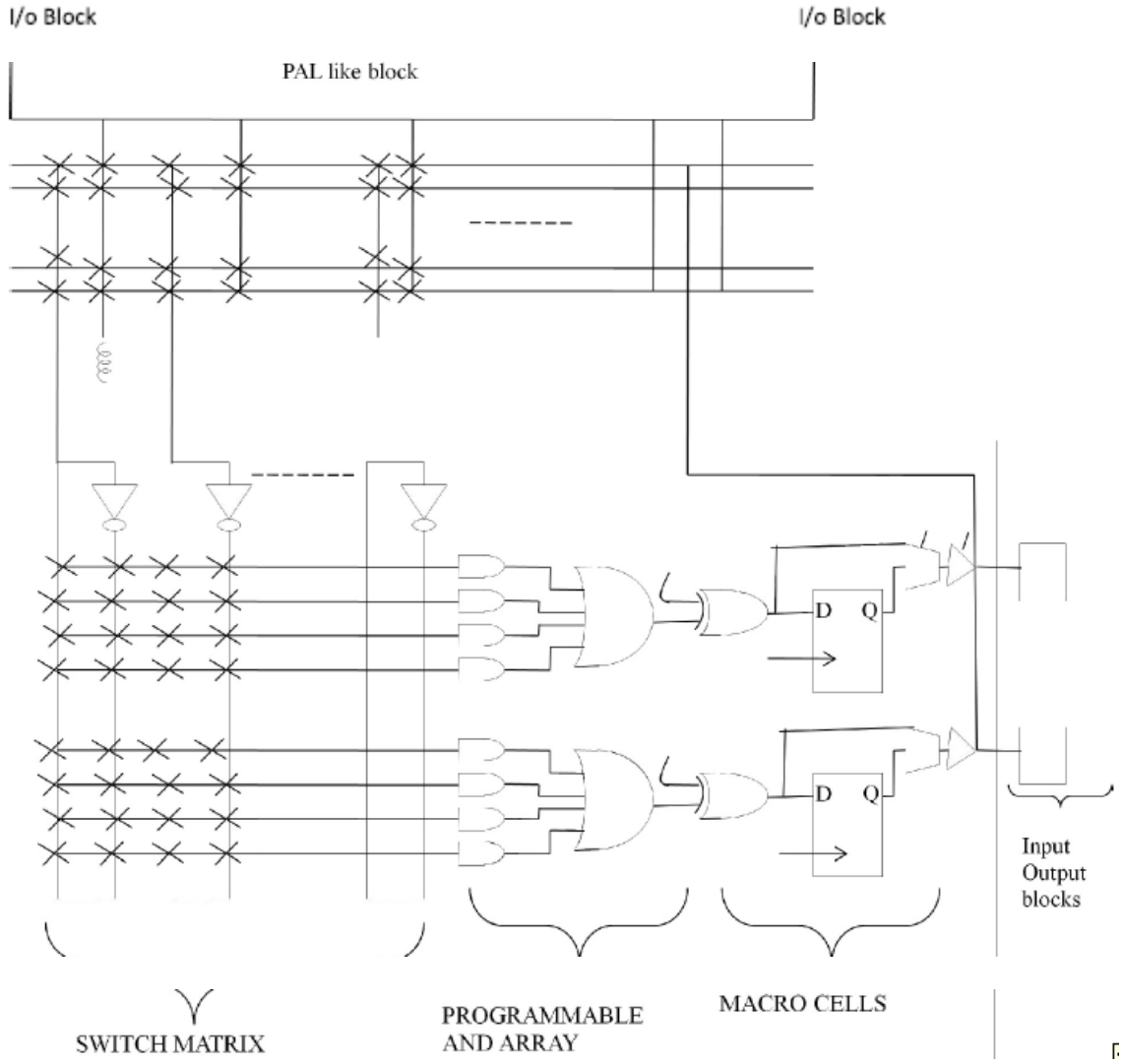
Complex Programmable Logic devices:

CPLD is a collection of individual PLDS on a single chip, accompanied by a programmable interconnection structure that allows other. PLDS to be looked up to each other. Here the chip area for n times as much Logic is only n times as much Logic is only n times the area of a single PLP plus the area of the programmable interconnect structure

General CPLD Architecture

The figure includes four PAL Like locks that are connected to a set of inter connections wires. Each PAL like lock is connected to I/O lock, which is attached to chips input and output pins.


I/o Block
I/o Block
PAL / PLD block
PAL / PLD block
Interconnection wires
PAL
BLOCK
PAL
BLOCK


I/o Block
I/o Block
PAL like block
D Q
D Q
Input
Output
blocks
SWITCH MATRIX
PROGRAMMABLE
AND ARRAY
MACRO CELLS

The figure shows the tactical wiring stretchers and connections in a PAL lock in a CPLD. The PAL Like lock includes two macro cells, each consisting of four input OR gate. The gate is connected to the nor gate. One input of the XOR gate can be Programmable connected to 1 or 0 – if 1 XOR gate complements OR gate o/p and if O then nor gate has no effect.

The flip flop is used to store the o/p value produced by OR gate. Each tri-state buffer is connected to a pin on the CPLD Package. When tri-state buffer is enabled, the pin is used as an o/p pin. When it is disabled the pin is used as an i/p pin.

The interconnection wiring contains programmable switches that are used to connect the PAL Like blocks.

Commercial CPLDS range in size from only 2 PAL Like block to more than 100 PAL – Like blocks. CPLDS are available in a variety of packages such as

PLCC – Plastic Leaded Chip Carrier

QFP – Quad flat package.

STANDARD CELLS:

A cell based ASIC, commonly termed as CBIC, uses predesigned logic cells Known as STANDARD CELLS. Example of these Predesigned cells include : AND gates, OR gates, multiplexers and flip flops.

Standard cell methodology is a method of designing (ASICS) with mostly digital Logic features. It is an example of design alert action whereby a low level VLSI Layout is encapsulated into an abstract Logic representation. Cell – bared methodology maker it possible for one designer to focus on the high – Level aspect of digital design, while another designer toward on the implementation of Physical aspect.

A standard Cell is a group of transistor and interconnect structures, which provides a Boolean Logic function.

Usually the initial design of a standard cell is developed at the transistor Level, in the form of a transistor net list. The net list is a nodal description of transistors, of their connections to each other, and their terminals (Ports) to the external 2 environment. Designers use CAD Programs such as SPICE to simulate the electronic behavior of net list, by declaring input stimulus and then calculating the circuits time domain response. The simulations verity whether the net list implements the requested function, and predict other pertinent parameters such as power consumption or signal Propagation delay.

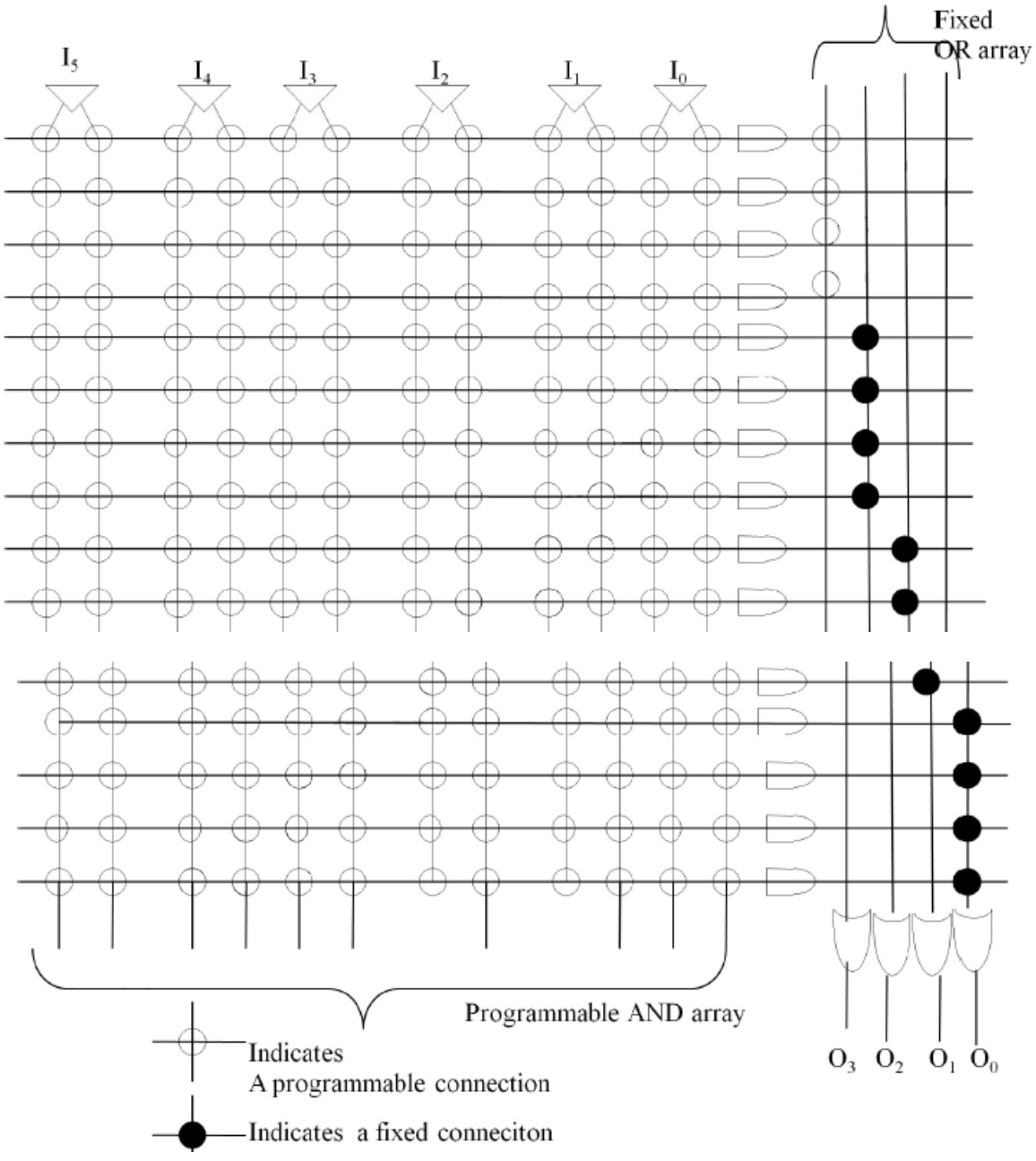
Since the Logical and net list views are only useful to abstract (algebraic) simulation, and not device fabrication, the physical representation of the standard cell must be designed too. Also called the Layout view, this is the lowest level of design abstraction in common design Practice.

Application of Standard cell:

A 2 input NAND or NOR function is sufficient to form any arbitrary Boolean function set. But in modern ASIC design, standard cell methodology is practiced with a Sizeable Library of cells. The Library usually contains multiple implementations of the same Logic function, differing in area and speed.

PROGRAMMABLE ARRAY LOGIC:

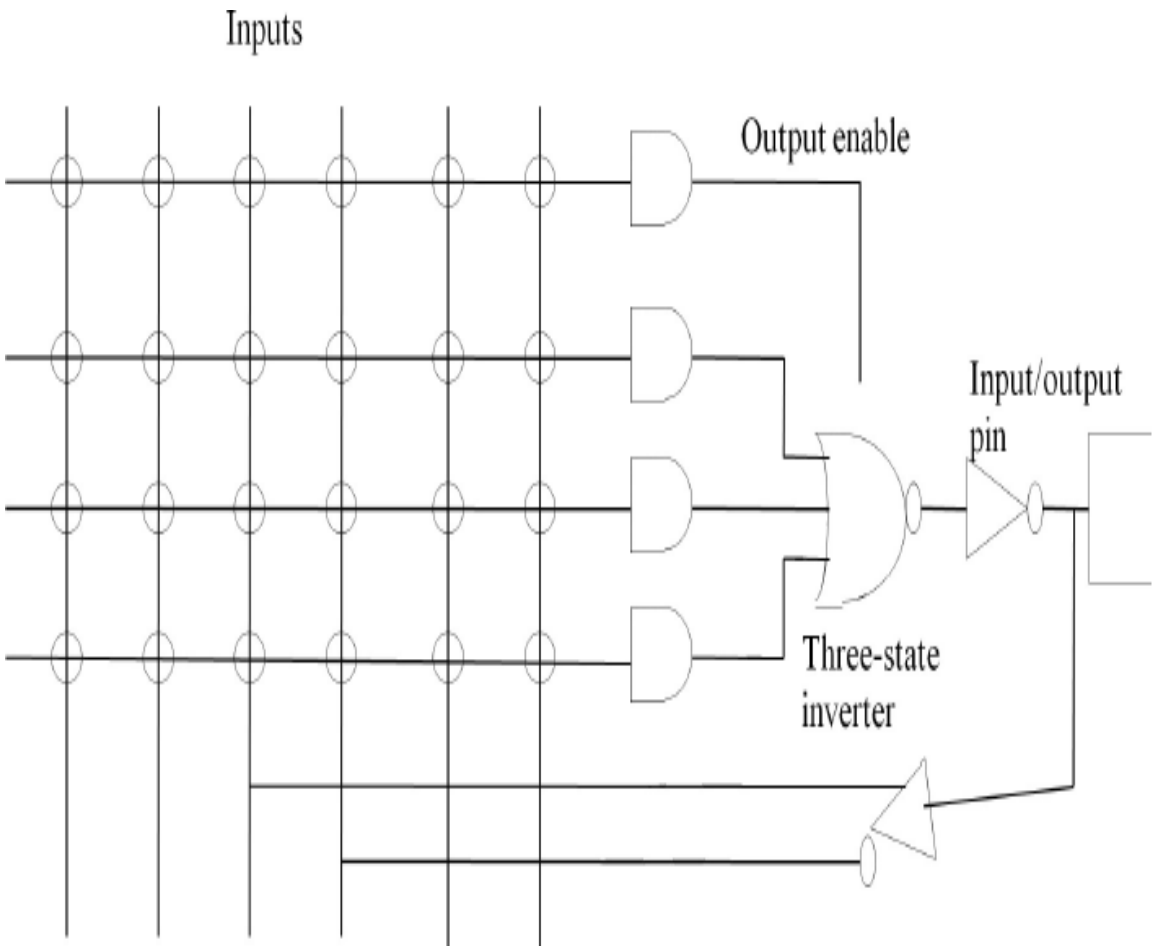
PAL has a less complex arrangement than PLA.



The figure shows a PAL Structures with six inputs, four outputs and 16 product terms. The AND array is programmable to select the min terms required for a given function. The OR array is fixed.

To give increased flexibility many PALS are a technique that Permits some of their Pins to be used either as inputs or as outputs.

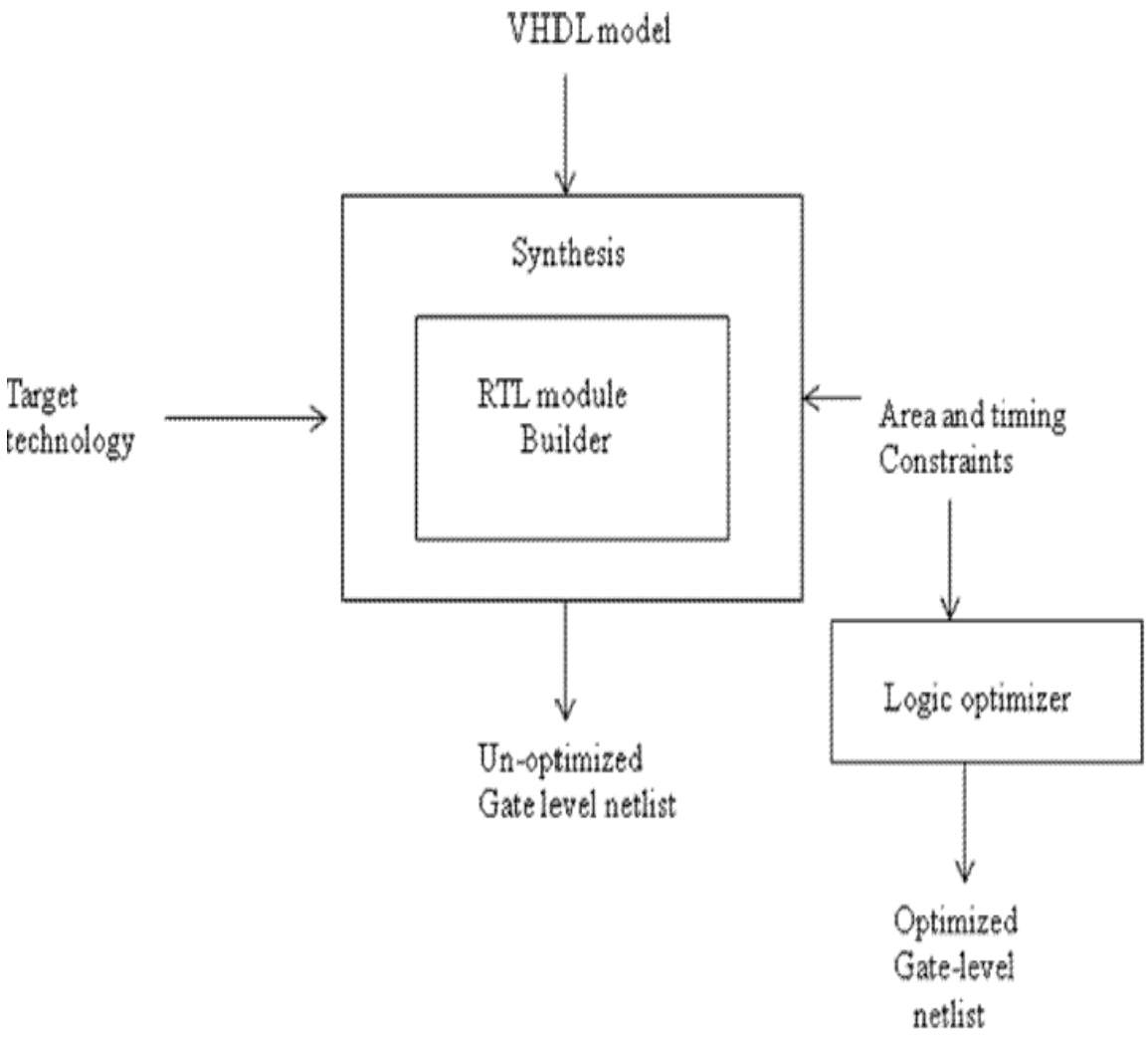
The following figure shows the arrangement.



Here the output from one of the OR gates of the device is passed through a three – state inverter before being fed to the output Pin. The operation of the inverters controlled by an output enable signal that is derived from the AND array in a manner similar to any other minterm. If all the fuses connected to the inputs to this AND gate is blown, its output will remain high, enabling the output of the inverter. This will configure the Line as an output.

VHDL SYNTHESIS:

It is the process of constructing a gate level net list from a model of a circuit described in VHDL. The process is explained by the following figure.



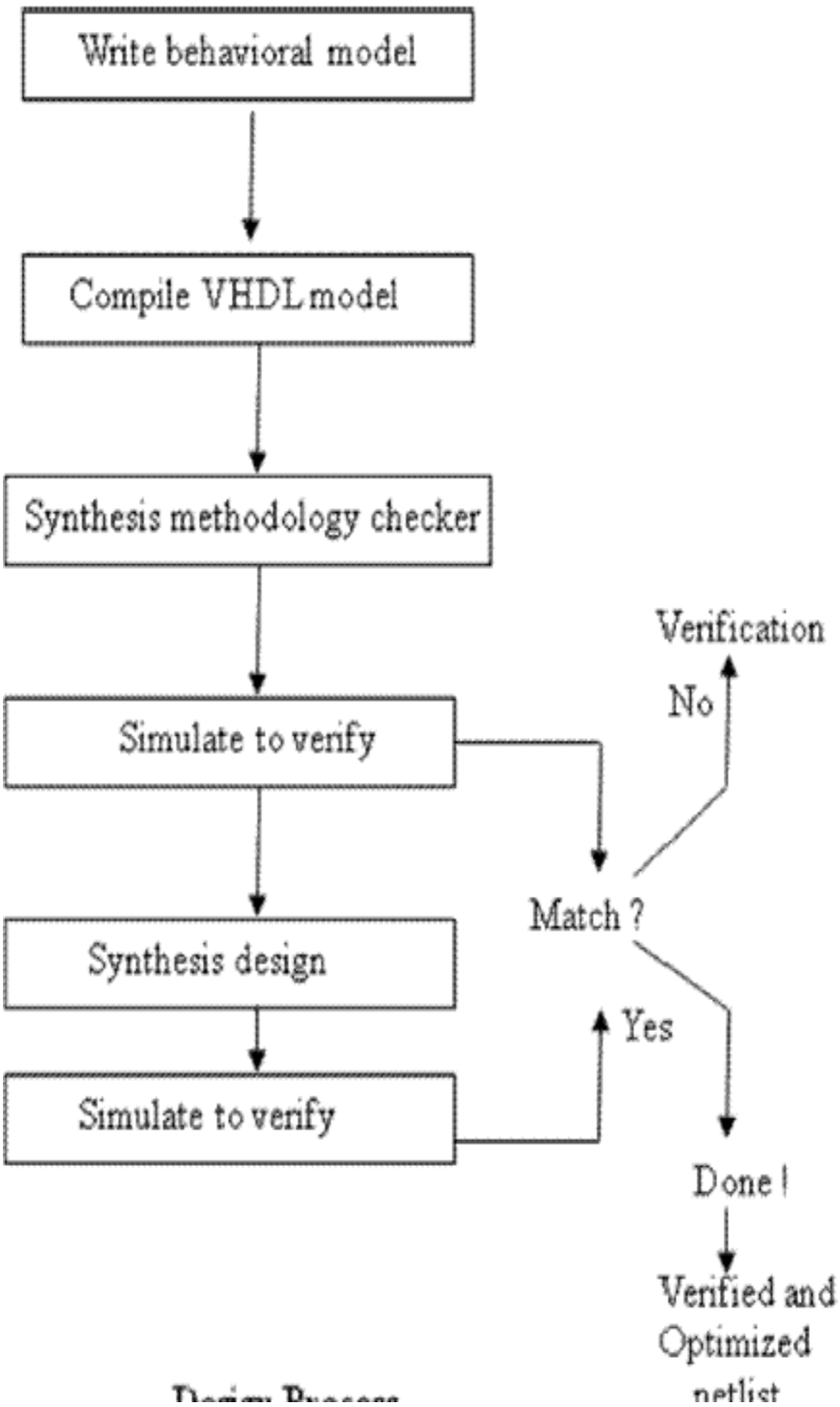
A synthesis process may alternately a RTL net list, which is comprised of generate Register – Transfer level blocks.

Having produced a gate-level net list, a logic optimizer reads this net list and the circuit for the user specified area is optimized. These constraints are also used by the module builder for appropriate selection or generation of RTL blocks.

With the help of VHDL the designer can model a circuit at different levels of abstraction ranging from the gate level, RTL lever, and behavioral level to the algorithmic level.

There is no standardized subset of VHDL, for synthesis. Each synthesis system may provide a different mechanism to model a flip-flop or a latch. Therefore, an own subset is defined by each system.

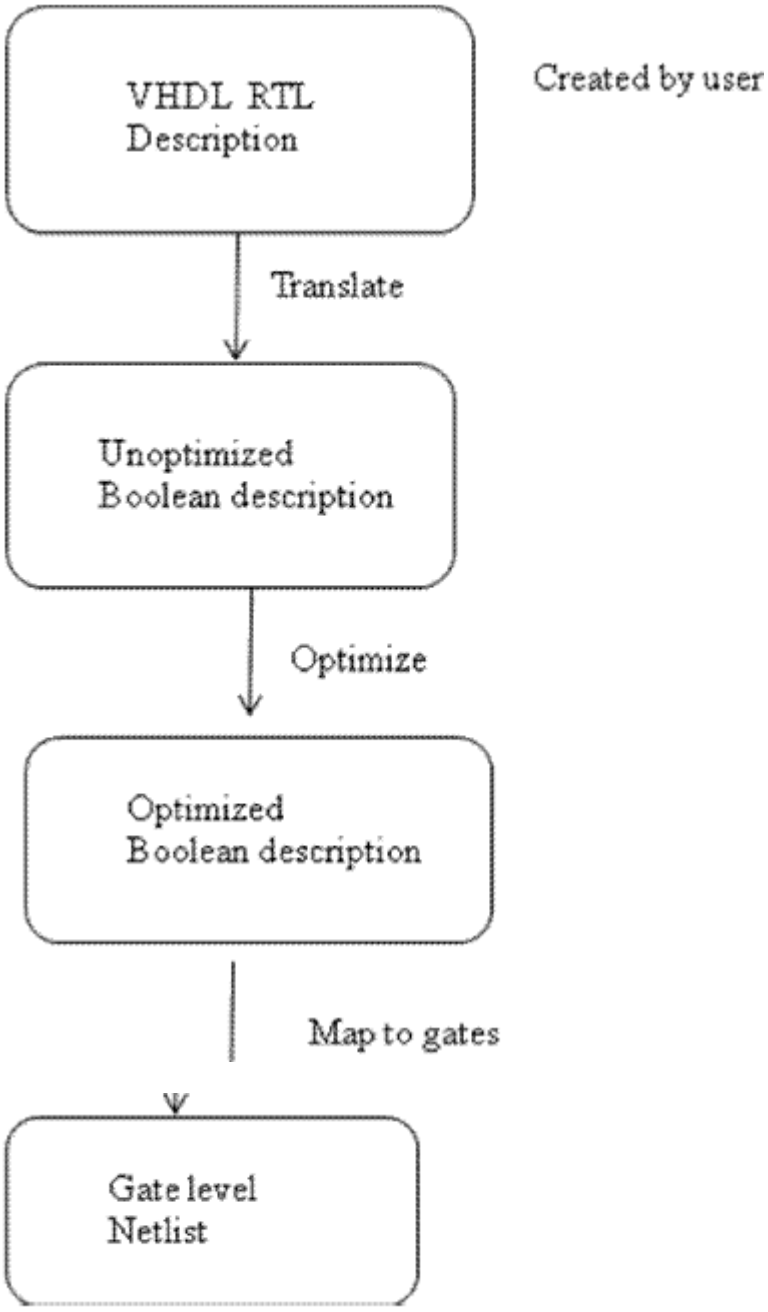
The design process shown in the following figure has to be followed for VHDL synthesis:



In this design process, a synthesis methodology checker is needed to check if the model being written is compatible for synthesis or not. This is done in the first Simulation loop. In this way, after the Simulation results have been verified, a verified synthesizable model exists, which can then be synthesized.

VHDL Synthesizer:

The RTL description is converted to gates by three steps basically.



The RTL description at first is converted to a un optimized Boolean description which consists of basic gates, flip flops and latches. This is functionally correct but completely an unoptimized description. Next, Boolean optimization algorithms are executed on this Boolean equivalent description to produce a optimized Boolean equivalent description. Finally this is mapped to acheal logic gates by using technology library of target process. Each stage is explained below:

Translation: It is done from RTL description to unoptimized Boolean description, not user controllable. All if, case and other statements are converted to their Boolean equivalent in this intermediate form.

Boolean optimixation: If converts the unoptimized Boolean description to optimized Boolean description. Many algorithms and rules are used to do this
Ex: convert to very low-level description (PLA format) and then optimized by reducing the logic generated by sharing common terms.

Flattening: The process of converting unoptimized Boolean description to PLA format is known as flattening, because it creates a flat signal representation of only two levels: an AND level and an OR level.
Ex: x= y and z
Where y = a or (b or c)
Z = c or d
Then x= (a or (b or c)) and (c or d)
Converting into a single statement removing the intermediate variables is called flattening.

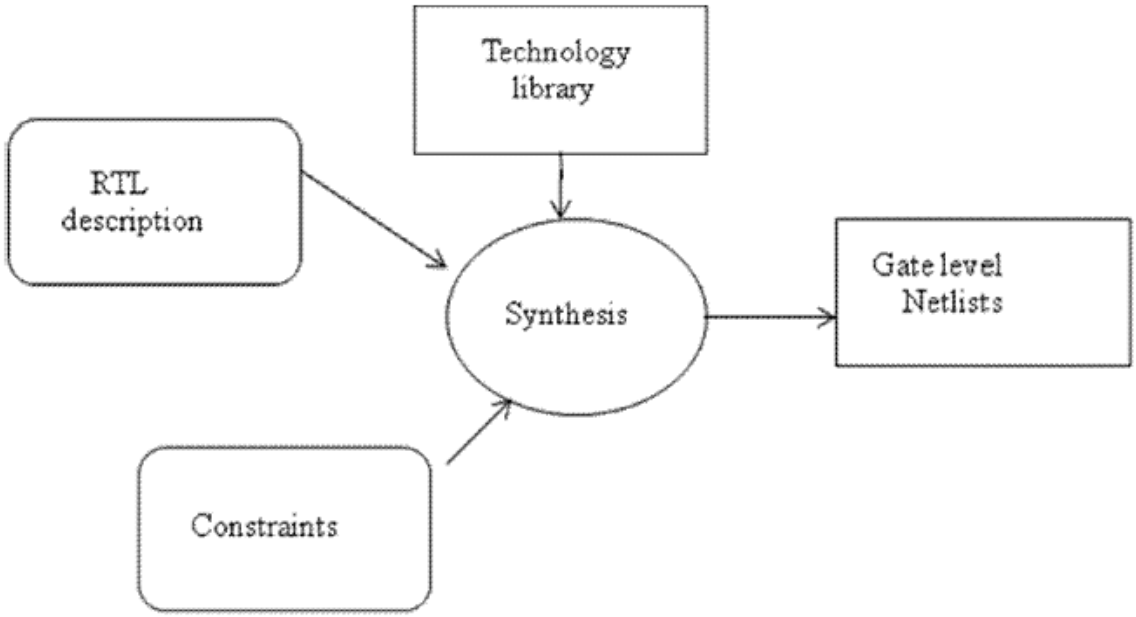
Factoring: It is the process of adding intermediate terms to a description to add structure tot it. It is there verse process of flattening.

Mapping to gates:
The logical and timing information from a technology library is taken by the mapping process to convert optimized Boolean description to a net list. There are number of net lists which are functionally same but differ in area and speed.

CIRCUIT SYNTHESIS AND DESIGNFLOW:

Circuit synthesis provides a path between VHDL and a netlist, analogous to the c compiler which provider path between C code and machine langeg. The circuit synthesis methodology is needed to check if the model being written is compatible for synthesis or not. The RTL descriptions are converted to the gate lee el netlist by the current synthesis tools available. Interconnected gate level macro cells make up the gate level netlist. The technology libraries consist the models for gate level cells. These nelilsts are optimized for area, speed, testability and so on.

The synthesis process is shown in the following figure:

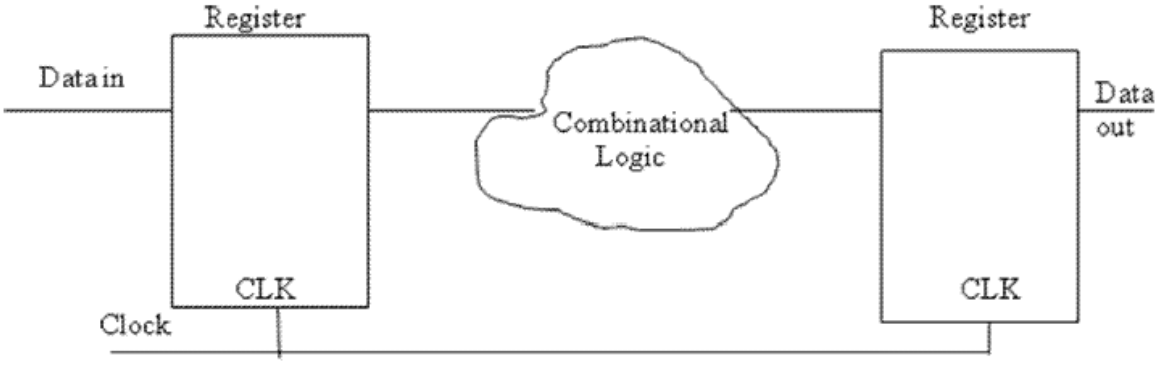


This process has the following as inputs:

1. RTL VHDL description
2. Circuit constraints
3. Attributes for the design and
4. Technology Library

1. RTL VHDL description:

A register transfer level description is characterized by a style that specifies all of the registers in a design and the combinational logic between them. This is shown by the register and cloud diagram in the figure below:

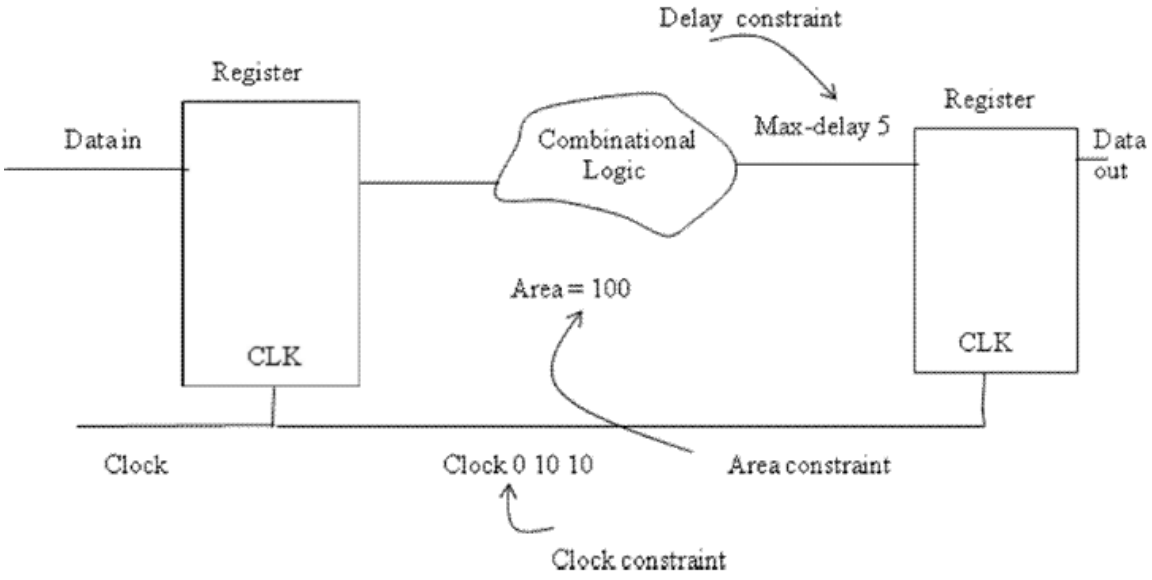


Register and Cloud diagram

The registers are described through component instantiation or inference; RTL descriptions are used for synchronous designs and describe the clock by clock behavior of the design.

2. Constraints:

Constraints are used to control the output of the optimization and mapping process. They provide goals that the optimization and mapping processes try to meet and control the structural implementation of the design. Constraints include area, timing, power and testability constraints. The most common are timing constraints. The block diagram is shown below:



There are different types of constraints as explained below:

(i) Timing constraints:

- These are used to specify maximum delays for particular pathes in a design.
- A typical delay constraint is shown here:

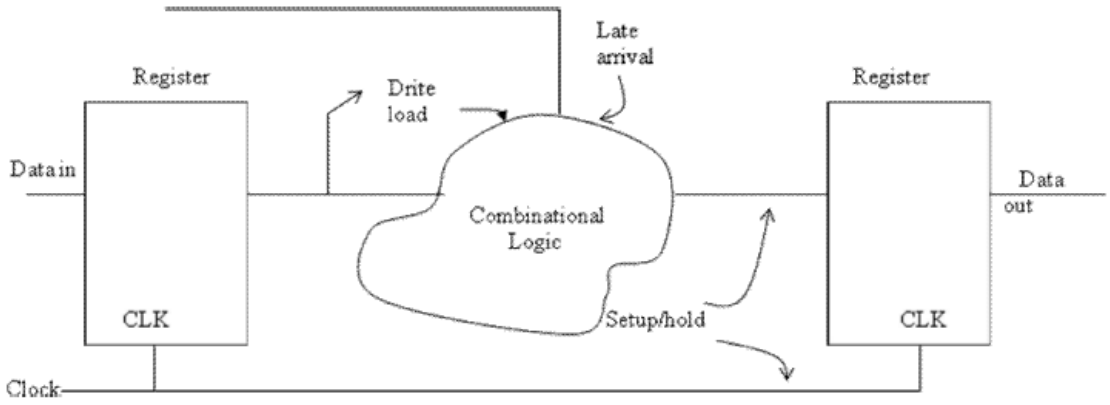
  Set-attribute-port data-out-name, required-time-value 25
- This constraint specifies that the maximum delay for signal data-out should be less than or equal to 25 library units.

(ii) Clock Constraints:

- We add a required time-constraint to every flip-flop input with the value of clock cycle.
- Another method is to add a clock constraint to the design.
- An example clock constraint is set – attribute – port clk- name clock- cycle – value 25

3. Attributes:

These are used to specify the design environment. For example, loading the output devices hoe to drive, the drive capability of devices driving the design and timing of input signals are specified by the attributes. A cloud diagram is shown in figure below:



Load:

The name of loads that can be driven within a particular time is determined by the drive capability specified by each output. Each input can have a load value specified, that determines how much it will show a particular driver.

The load attribute specifies how much capacitive load exists on a particular output signals.

Ex: Set- attribute-port x bus-name input-load- value 5

Drive:

The drive attribute specifies the resistance of the driver, which controls how much current it can source. An example of a drive specification is:

Set-attribute-port y bus- name output- drive- value 2.7

The attribute specifies that signal y bus has 2.7 library units of drive capability.

Arrival Time:

Setting the arrival time on a particular node specifies to the static timing analyzer when a particular signal will occur at a node.

Late arriving signals drive inputs current block at a later time, but the results of the current block still must meet its own timing constraints on its outputs.

4. Technology Library:

Technology Libraries hold all the information necessary for a synthesis tool to create a netlist for a design based on the desired logical behavior, and constraints on the design.

These libraries contain all the information that allows the synthesis process to make the correct choice and build a design. They contain not only the logical functions of an ASIC cell, but the area o the cell, the input to output timing of the cell, any constraints on fanout of the cell, and the timing checks that are required for the cell.

SIMULATION:

A Simulation is used to predict and verify the performance of a given circuit. They can be divided into following categories:

1. Transistor-level or circuit-level Simulation
2. Static timing analysis or timing Simulation
3. Gate level simulating
4. Switch level Simulation
5. Behavioural Simulation
6. Functional Simulation.

1. Circuit-level Simulation:

It is the most accurate technique. These Simulations operate at the circuit level. It is complex and time consuming. Here the electrical behavior the various parts of circuit is to be implemented inn silicon. Circuit analysis programs are typified by SPICE program. Simulation time is proportional to N^m.

Simulation time α N^m

N. number of non linear devices in the circuit

m= varies between 1 to 2

Timing Simulation:

Once a behavioral or functional Simulation predicts that a system works correctly, the next step is to check the timing performance. At this point a system is partitioned into ASICS and a timing Simulation is performed for each ASIC separately. Timing analysis in a static manner computing delay times for each path is called static timing analysis. The path with the longest delay is called critical path.

The structure of the timing tools ensures that the run times are strictly linearly related to the number of devices and nodes being simulated.

Logic level Simulation (or) gate-level Simulation:

Here the performance is assessed in terms of logic levels with no or little timing information. They can simulate large section of layout at one time.

In a gate level Simulation a logic gate or logic cell is located as a black ox modeled by a function hose variables are the input signal.

These Simulations use primitive models such as NOT, AND, OR, NAND and NOR gates. They can be operated in unit delay mode or timing parameters may be assigned based on prior circuit Simulation, and known circuit parasitic.

Timing is normally specified in terms of inertial delay and load dependent delay for the appropriate edge transition, as follows:

$$T_{gate} = T_{inteinsic} + C_{load} \times T_{load}$$

(delay of gate)	(Intrinsic Delay of Gate)	(actual Load in Some units) i.E PF	(delay per Load)

Logic Simulations with such timing information are accurate for Cmos logic configurations.

Switch level Simulation:

This type of Simulation models transistors as switches – on or off. It can provide more accurate timing predictions than gate-level Simulation, but without the ability to use logic cell delays.

They merge logic-Simulation technique with some circuit Simulation techniques by modeling transistors as switches.

Behavioral Simulation:

Large pieces of system modeled as block boxes with inputs and outputs crating an imaginary Simulation model of the system are called behavioral Simulation.

Functional Simulation:

Functional Simulation ignores timing and includes unit-delay Simulation, which sets delays to a fixed value 7.31.

In the processes in ASIC design flow there are two kinds of Simulation

a. Post layout timing Simulation
b. Post synthesis Simulation.

a. Post layout timing Simulation:

After place and route process has completed the designer uses post route gate level Simulation to verify its results.

This Simulation combines the netlist used for place and route process into a Simulation that checks both functionality and timing of the design. The designer can run the Simulation and generate accurate output waveforms that show whether the device is working properly and if the timing is being met.

Same test vectors and the same Simulation as for the RTL Simulation can be used for post route gate level Simulation if properly structured.

b. Post synthesis Simulation:

This Simulation is carried out after the synthesis has been done.

This is as effective as the post layout timing Simulation.

Applications of Simulation:

Simulation is applied in two major application areas:

- System validation
- Fault Simulation

1. System validation:

In system validation we to deduce that the system conforms to a specification. The system is represented by a model at different points in the design process. Based on the design cycle, a system can have models at the chip, register and circuit level. One uses Simulation to validate that these models conform to that specification.

2. Fault Simulation:

The other application of Simulation is fault Simulation. Here, faults are injected into the system model, and the system is simulated to observe the response. To establish that a test detects a fault this can be done.

This type of Simulation can also be used to create fault dictionaries, which relate output signal states to faults, therefore allowing one to diagnose which fault occurred.

Efficiency of a Simulation:

When simulating models for large systems Simulation efficiency is very important. For logic Simulation, Simulation efficiency (E) is defined as

$$E = \frac{\text{Real Logic time}}{\text{Host CPUtime}}$$

'Real logic time' is the actual time required to complete an activity sequence in real logic circuit. 'Host CPU time' is the time required to simulate the same activity sequence using a logic Simulation running on a host CPU.

Advantages of Simulation:

1. It permits evaluation of a part before it is available.
2. It permits substitution of a range of components into signal paths to study the effects of different timing.
3. The ability to see inside ICs, so that the designer can examine the contents of the registers and flip flops during simulation is one of the benefits.
4. The Simulation can also be used to detect glitches caused by hazards.
5. The Simulation can be designed to detect instances where two or more tristate devices are simultaneously active.

LAYOUT

Graphical Entry Layout:

Special textual entry editors are used for small subsystem layout. These tools have been replaced by highly interactive method of producing layouts for which monochrome or color graphics terminals are used. The layout is built up and displayed during the design process.

These systems are menu driven and possible actions at various stages of the design are displayed on the screen along with details of the current layout.
Two of the earliest available graphical entry packages were KTC and PLAN. PLAN makes use of low- cost monochrome as well as colour graphical terminals.

RC Calculation from layout:

It can be described in terms of sheet resistance and wiring capacitance.
The following table gives the sheet resistance values for different layers:

	Layer	Sheet resistance
1	Metal	0.03
2	Diffusion	10 → 50
3	Silicide	2 → 4
4	Polysilicon	15 → 100

For any layer, Area capacitance is

$$C = \frac{\epsilon_0 \epsilon_i \; ns.A}{D} \; farads$$

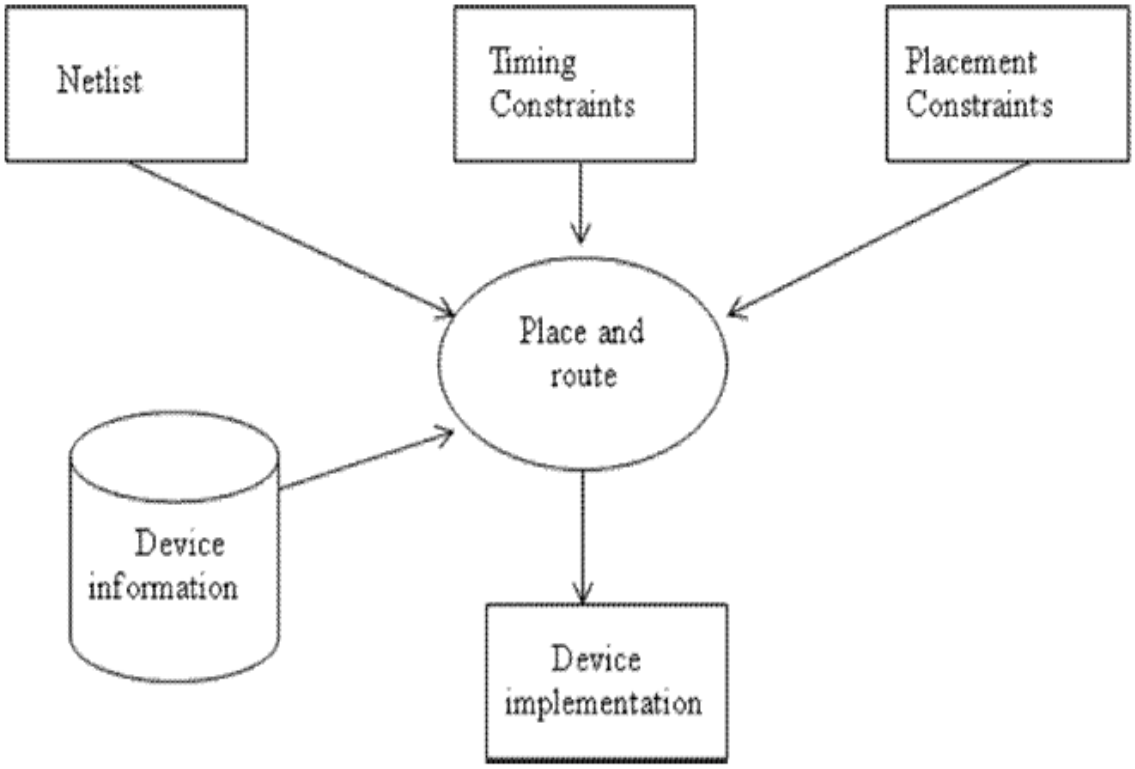
D= Thickness of Silicon dioxide
A= area of plates
ε_{ins}= Relative permittivity of SiO_2 =4
$\varepsilon_0 = 8.85 \times 10^{-14}$ p/cm.

DESIGN CAPTURE AND DESIGN VERIFICATION TOOLS:

Place and Route tool:

These tools are used to take the design netlist and implement the design in the target technology device. They place each primitive from netlist into an appropriate location on the target device and then route signals between the primitives to connect the devices according to the netlist. These tools are very architecture and device dependent. Place and route tools for FPGA and ASIC devices can be obtained from the respective vendors.
The data flow diagram of the place and route tools is shown below:

Inputs to the place and route tools are the netlist in EDIF or another netlist format, and possibly timing constraints.
Another input is the timing constraints which informs the tool about the signals which have critical timing associated with them, and to route these nets, into most timing efficient manner. These constraints tell the place and route tool to place the primitive in los proximity to one another and to sue the fastest routing.
Placement of large parts of the design is known as floor planning. It allows the user to pick location on the chip for large blocks of design so that routing wires are as shot as possible.
After all the cells are placed and routed, the output of the place and route tools consists of dataflow that can be used to implement the chip.



The other output from the place and route software is a file used to generate the timing file. This file describes the actual timing of the programmed FPGA deice or the final ASIC deice.

CAD Tools:

Building a chip requires a variety of CAD Tolls. CAD algorithms are careful chosen because of the complexity.
Most CAD design data consist of structural description graphs, geometric objects etc. Most programming tools appear to the user to be manipulating test files. Layout representation requires geometric information,. Such as rectangles and triangles. Circuit information is stored as enlists or component lists. The CAD includes.

1. Physical design layout and editing capabilities which may be through textual or graphical entry of information.
2. Structure generation/system composition consist of capabilities of design layout software of point 1.
3. Physical verification: These tools include design rule checking (DRC), circuit extractor ratio run and other static checks and to plot out and/or display for visual checking.
4. Behavior verification: Simulation at various levels will be required to check out the design before one embarks on the expense of turning out the design in silicon.

Design Rule Checkers:

All possible errors must be eliminated before mask making proceeds.

- Check for errors at all stages of design
- The different stages are:

- Pencil and paper stage of the design of leaf cells.
- At the leaf cell level/once the layout is complete.
- At the subsystem level check that butting together and wiring up of leaf cells is correctly done.
- Once the system layout has been completed.

Circuit Extractors:

If the design information exists in the form of physical layout then a circuit extractor program is required that wile interpret the physical layout in circuit term. It is fed directly into simulator that computer may be used to interpret the finding of the extractor.

Design verification Prior to Fabrication:

- It is not sufficient to have good design tools, there must be complemented by equally effective verification software capable of handling large systems and with reasonable computing power requirements.
- The nature of tools required will depends on the way in which an integrated circuit design is represented in computer.
- There are two approaches:

(I) Mask level layout languages such as CIF, which are well suited to physical layout description but not for capturing the design intent.

(II) Circuit description languages where the primitives are circuit elements such as transistors, wires and nodes.

TEST PRINCIPLES:

The responsibilities for the various levels of testing and testing methodology lie on the designer. Small imperfections in starting material, processing steps or in photo masking may result in bridged connection or missing features. The aim of a test procedure is to determine the good die to be used in the end system.
Testing a die can occur at

- ✓ Wafer level
- ✓ Packaged chip level
- ✓ Board level
- ✓ System level
- ✓ Field

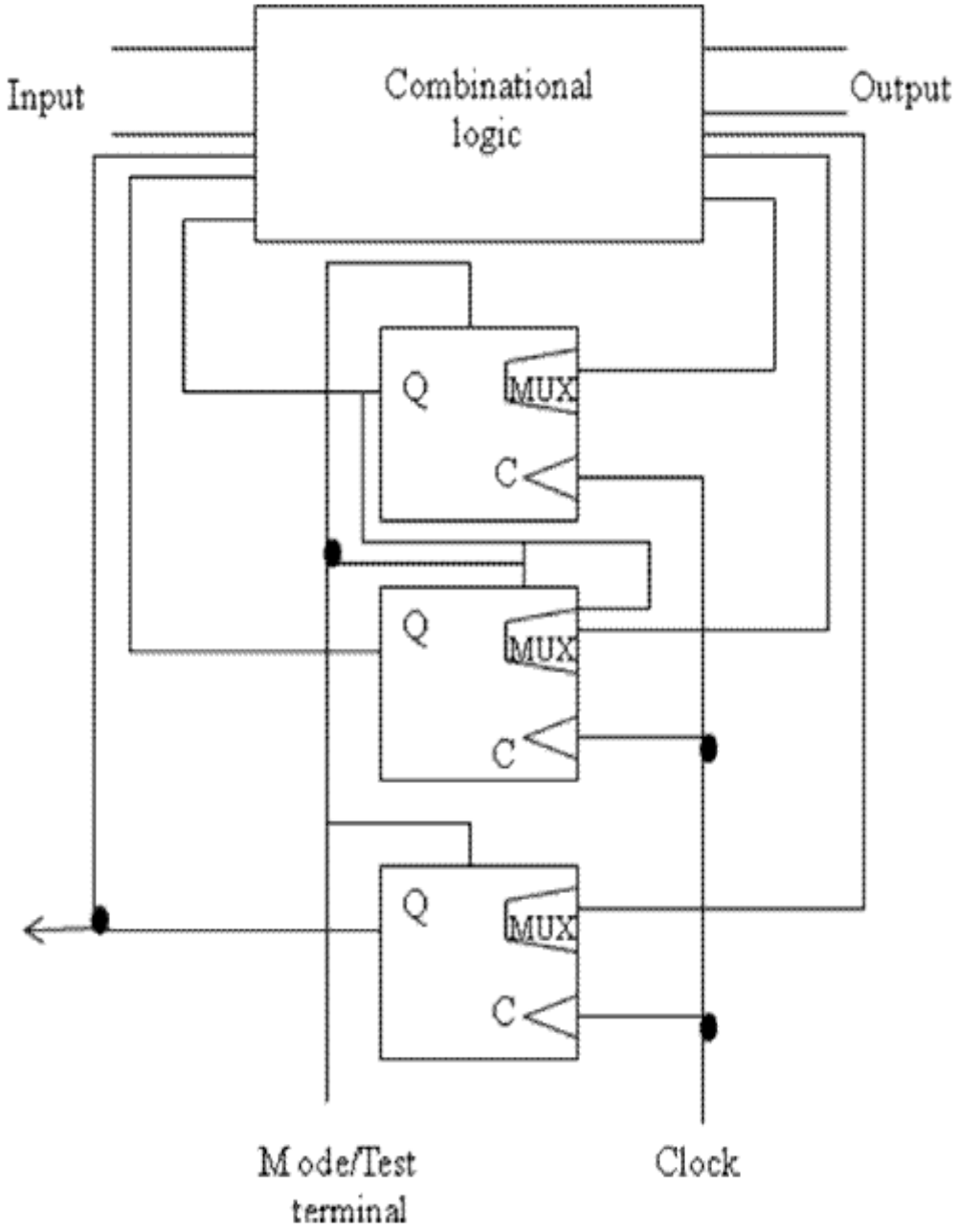
If we detect a malfunctioning at a earlier stage the cost of manufacturing is minimum. Some of the test principles are:

(i) Scan – Based Techniques
(ii) Boundary scan test
(iii)Built-in-self test techniques.

(i) Scan – Based Techniques

If more accessible logic nodes with use of additional primary input lines and multiplexers, controllability and observability can be enhanced. But this can be costly, so an alternative s to use scans registers with both shift and parallel load capabilities. The scan design technique is a structured approach to design sequential circuits for testability. The storage cells in registers are used s observation points, control points, or both. The testing of a sequential circuit is reduced to the problem of testing a combinational circuit.

In general a sequential circuit consists of a combinational circuit and some storage elements. The storage elements are connected to from a long serial shift register, so called scan path, by using multiplexers.
In test mode, the scan-in-signal is clocked into the scan path and the output of the last stage of latch is scanned out. The testing sequence as follows:
Step 1: Set the mode to test and let latches accept data from scan-in-input
Step 2: Verify the scan path by shifting in and out the test data.
Step 3: Scan in the desired state vector into the shift register.
Step 4: Apply the test pattern to the primary input pins.
Step 5: Set the mode to normal and observe the primary outputs of the circuit after sufficient time for propagation.
Step 6: Assert the circuit clock for one machine cycle to capture the outputs of the combinational logic into the registers.

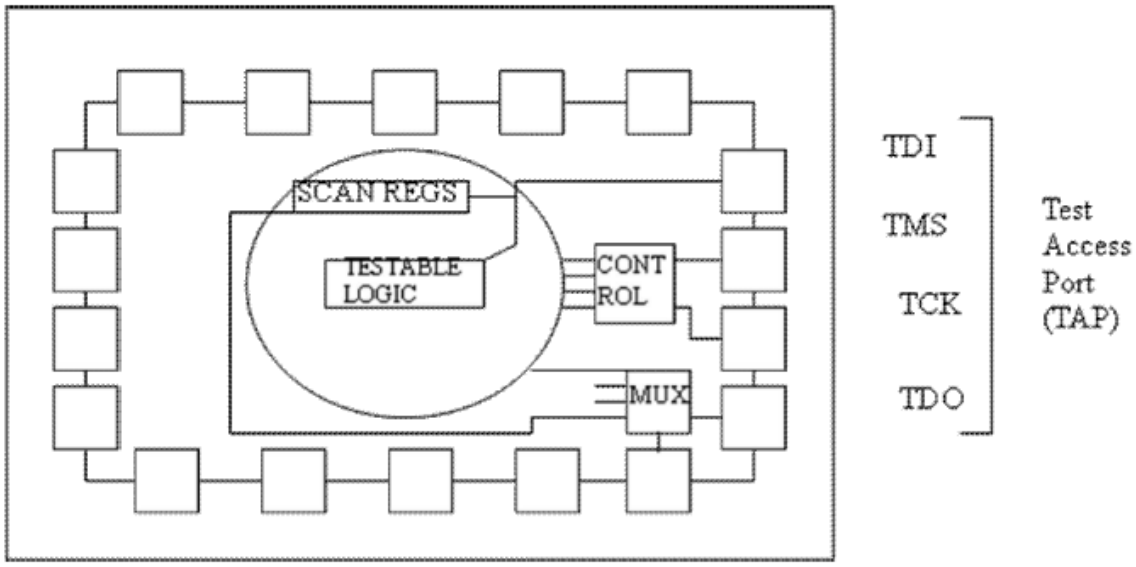


Step 7: Return to test made, scan-out the contents of the registers, and at the same time scan-in next pattern.

Step 8: Repeat steps 3-7 until all test patterns are applied.

(ii) Boundary Scan Test:

The problems associated with the testing of boards carrying VLSI circuits and /or surface-mounted devices are resolved by technique involving scan path and self-testing. It consists of placing a scan path (shift register) cell adjacent to each component pin and to inter connect the cells so as to form a chain around the border of circuit. The technique is explained in diagram below:



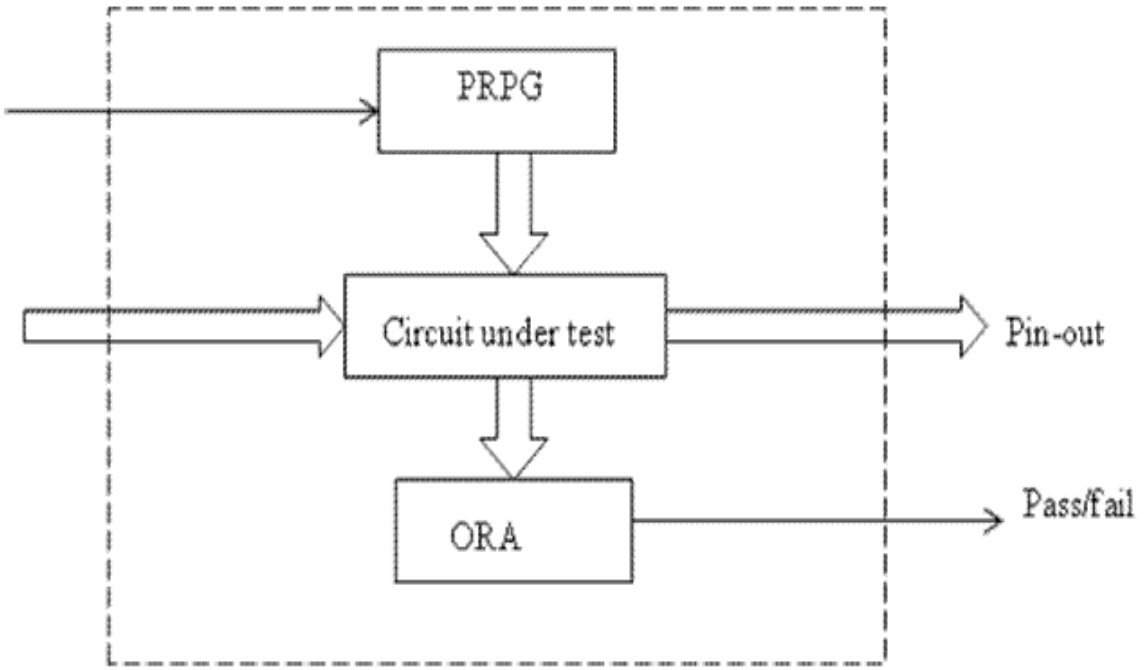
The boundary scan path is provided serial input and output pads and appropriate clock pole which make it possible to

- Test the inter connections between various chips
- Deliver test data to the chips on the board of self testing
- Test the chips themselves with internal self-test facilities.

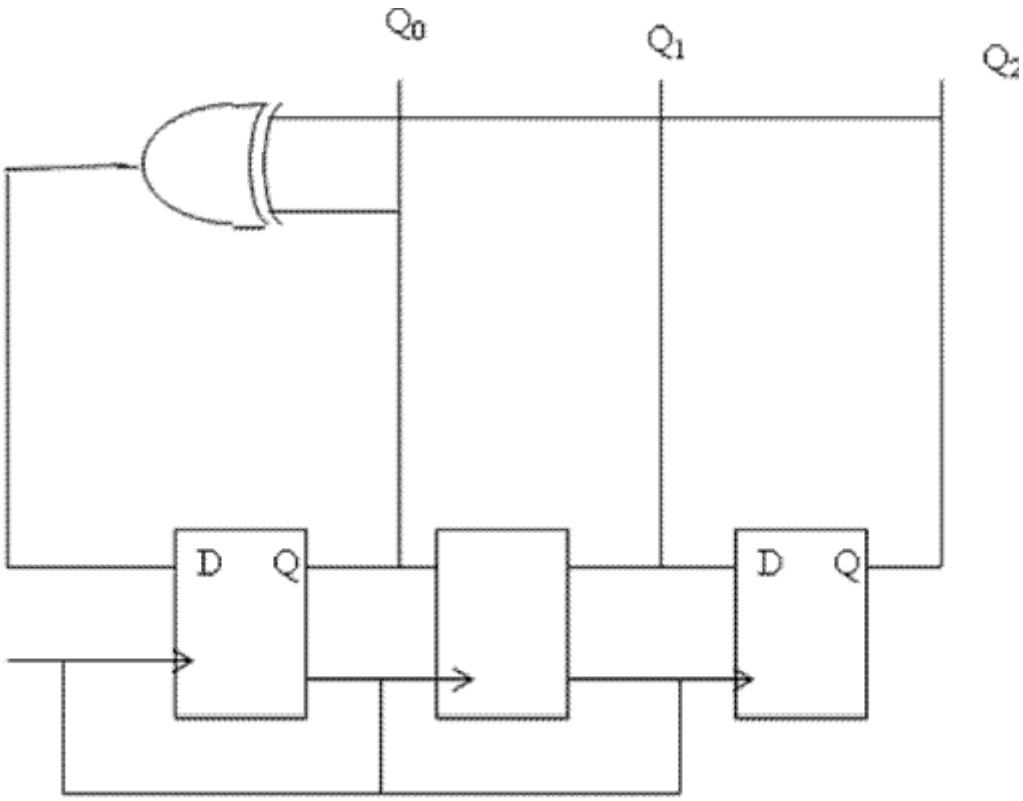
(iii)Built in self test techniques:

Here parts of the circuit are used to test the circuit itself online BIST is used to perform the test offline. The essential modules include.

- Pseudo random pattern generator (PRPG).
- Output Response analyzer (ORA).



The roles of there two modules are illustrated above. The implementation of both PRPG and DRA can be done with linear feedback shift registers (LFs Rs)
The patterns are to generated to test the circuit. This is done by pseudo random pattern generator. Its circuit is shown below:



A Pseudo-random sequence generator using LFSR

Design for Test (DFT):

There techniques are increasingly gaining momentum as they provide measures to comprehensively test the manufactured device for quality and coverage.
Merging testability features early in the cycle was the final solution, creating thename Design for Test.

Types of DFT:

The main DFT techniques available today are;

(a) Scan insertion – involves various design issues to be resolved

- Involves replacing all flip flops
- Most prevalent is the multiplexed flip flop.

(b) Memory BIST insertion:

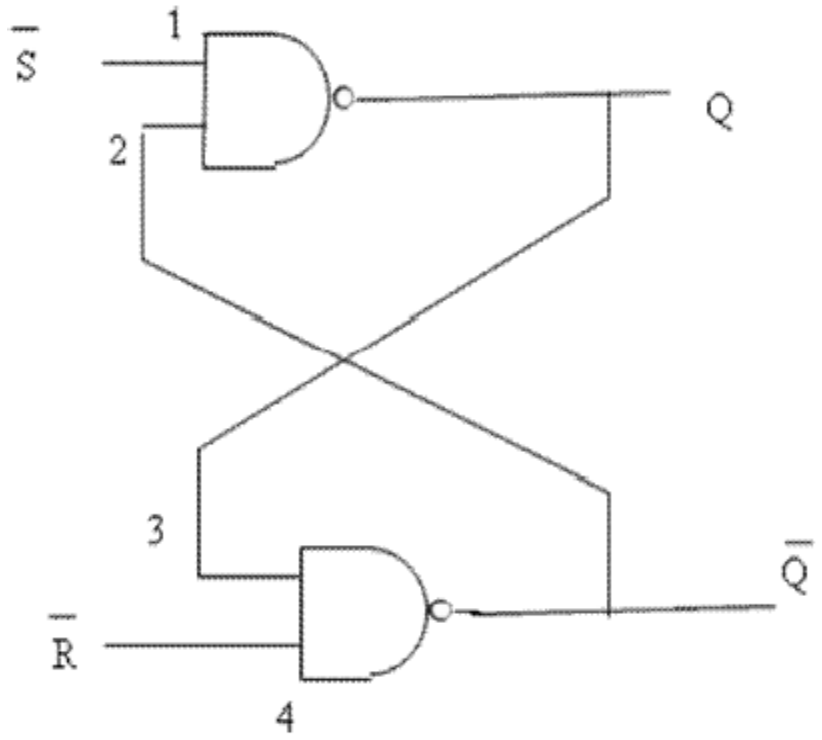
- It is comprised of a controller logic that uses various algorithms to generate input patterns that are used to exercise the memory elements of a design.
- It is automatically generated, based on the size and configuration of memory element.

(c) Boundary Scan DFT:

- Used for testing without unplugging the clip from the board.
- The TAG controller an surrounding logic also may be generated.

Procedure to test sequential logic:

All the sequential circuits exhibit a memory, property of remembering or taking into account the previous conditions. Consider the basic sequential circuit as shown below:



The sequential logic testing effects of memory.

The delay in the fed back path is non-existent, which is a case when the circuit propagation delays contribute to the necessary delay elements. The inert is applied to $\overline{S}, \overline{R}$.

The state of 'Q' for a "good machine (GM)" is tabulated first by the remaining columns and then for the' faulty machine' (FM) for a stud at 1 fault on each of the four inputs (1, 2, 3, 4).

The machine matches the good response in the first table with the SA 1 fault on the line 2 and so this particular test sequence will not detect a SA 1 fault on line 2.

The vectors are applied exactly in the reverse order in the second table. Each of the "?" indicated that the Q will retain the value it had earlier to the application of the test vector for that row. Again if the latches are reset i.e, Q=0 prior to application of the test sequence, then the SA1 fault on line 2 may not be detected.

Inputs		Output Q	
		GM	Input SAS 1
S	R		
			1 2 3 4
0	0	1	0 1 1 1
0	1	1	0 1 1 1
1	0	0	0 0 0 1
1	1	0	0 0 1 1

Inputs	Output Q

		GM	Input SAS 1
S	R		
			1 2 3 4
1	1	?	? 0 1 ?
1	0	0	0 0 0 ?
0	1	1	0 1 1 1
0	0	1	0 1 1 1

UNIT WISE IMPORTANT QUESTIONS

UNIT-I

1. Describe the two commonly used methods for obtaining integrated capacitor.
2. With neat sketches, explain in detail, all the steps involved in electron lithography process.
3. What is Moore's law? Explain its relevance with respect to evolution of IC Technology.
4. With neat sketches explain the fabrication of CMOS inverter using p-well process.
5. Explain in detail about NMOS enhancement mode of operation.
6. Explain various regions of CMOS inverter transfer characteristics.
7. Write in detail about integrated passive components.
8. Explain the MOS Transistor operation with the help of neat sketches in the following modes

   (a) Enhancement mode

   (b) Depletion mode.

9. Explain latch up problem in CMOS circuits.
10. (a)What are different VLSI technologies available compare their speed/power performance.

    (b) Why is VLSI design process presented in NMOS only?

    (c) Discuss the micro electronics evolution.
11. Draw the cross sectional view of CMOS P - Well inverter.
12. With neat sketches explain the NMOS fabrication procedure.
13. With neat sketches explain BICMOS fabrication process in an N well.
14. With neat sketches necessary, explain the oxidation process in the IC fabrication process.
15. Draw the basic design flow through typical CMOS VLSI tools and give some names of corresponding tools.
16. For a CMOS inverter, calculate the shift in the transfer characteristic curve When β_n/β_p ratio is varied from 1/1 to 10/1.
17. Explain different forms of pull ups used as load, in CMOS and in enhancement & depletion modes of NMOS.
18. Determine the pull up to pull down ratio of an nMOS inverter driven by another nMOS transistor Explain nMOS inverter and latch up in CMOS circuits?

19. Derive an equation for I_{DS} of an n-channel Enhancement MOSFET operating in Saturation region.
20. An nMOS transistor is operating in saturation region with the following parameters.V_{GS} = 5V; V_{tn} = 1.2V ; W/L = 110; $\mu_n C_{ox}$ = 110 μA/V ^{2}. Find Transconductance of the device.
21. (a) Derive an equation for Transconductance of an n channel enhancement MOSFET operating in active region.

    (b) A PMOS transistor is operated in triode region with the following parameters.V_{GS}=- 4.5V, V_{tp}= -1V; V_{DS}=-2.2 V, (W/L) =95, $\mu_n C_{ox}$ =95μA/V ^{2}. Find its drain current and drain source resistance.

UNIT-II

1. What are design rules? Why is metal- metal spacing larger than poly –poly spacing.
2. (a) What is a stick diagram? Draw the stick diagram and layout for a CMOS inverter.
    a. (b) What are the effects of scaling on V_t?
3. Draw the stick diagram and mask layout for a CMOS two input NOR gate and Stick diagram of two input NAND gate.
4. Draw the stick diagram and a translated mask layout for nMOS inverter circuit.
5. Explain the following
    a. Double metal MOS process rules.
    b. Design rules for P- well CMOS process.
6. Design a stick diagram and layout for two inputs CMOS NAND gate indicating all the regions & layers.
7. (a) Discuss design rule for wires (orbit 2μm CMOS).

    (b) Discuss the transistor related design rule (orbit 2μm CMOS).
8. Design a stick diagram and layout for the NMOS logic shown below $Y = ((A + B)C)^1$.

UNIT-III

1. Describe three sources of wiring capacitances. Explain the effect of wiring capacitance on the performance of a VLSI circuit.
2. Define and explain the following:
        i. Sheet resistance concept applied to MOS transistors and inverters.

ii. Standard unit of capacitance.

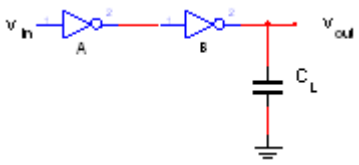
3. Explain the requirement and functioning of a delay unit.
4. Explain the requirement and operation of pass transistors and transmission gates.
5. Compare pseudo-n MOS logic and clocked CMOS logic.
6. Two nMOS inverters are cascaded to drive a capacitive load $C_L=14C_g$ as shown in Figure. Calculate the pair delay V_{in} to V_{out} in terms of τ for the given data.

   Inverter -A

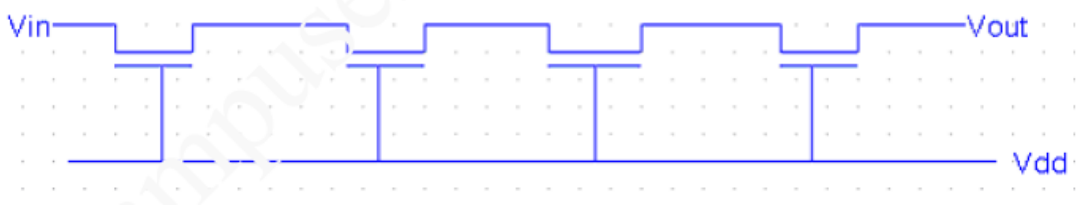
   $L_{P.U}= 12\lambda$, $W_{P.U} = 4\ \lambda$, $L_{P.d} = 1\ \lambda$, $W_{P.d} = 1\ \lambda$

   Inverter -B

   $L_{P.U}= 4\lambda$, $W_{P.U} = 4\ \lambda$, $L_{P.d} = 2\ \lambda$, $W_{P.d} = 8\ \lambda$



7. Determine an equation for the propagation delay from input to output of the pass transistor chain shown in figure 4a with the help of its equivalent circuit.



8. (a) Explain clocked CMOS logic, domino logic and n-p CMOS logic.

   (b) In gate logic, compare the geometry aspects between two -input NMOS NAND and CMOS NAND gates.

UNIT-IV

1. What are super Buffers?
2. Explain how a Booth recoded multiplier reduces the number of adders.
3. Draw the Schematic and mask layout of array adder used in Booth Multiplier and explain the principle of multiplication in Booth Multiplier.
4. Design a magnitude comparator based on the data path operators.
5. Draw the circuit diagram for 4-by-4 barrel shifter using complementary transmission gates and explain its shifting operation.
6. (a) How can the components of CMOS system design be categorized into the groups.

(b) Why is the static 6 transistor cell used for average CMOS system design?

(c) Compare the performance of CMOS Off chip and On chip memory designs.

9. Explain briefly the CMOS system design based on the data path operators, memory elements, control structures and I/O cells with suitable examples.
10. Draw circuit diagram of a one transistor with transistor capacitor dynamic RAM and also draw its layout.
11. Explain the tradeoffs between open, closed, and twisted bit lines in a dynamic RAM array.
12. Draw the typical standard-cell structure showing regular-power cell and explain it.

UNIT-V

1. Explain briefly the CMOS system design based on the data path operators, memory Elements, control structures and I/O cells with suitable examples.
2. Draw and explain the FPGA chip architecture.
3. Draw and explain the AND/NOR representation of PLA.
4. Draw the typical architecture of PAL and explain the operation of it.
5. What is CPLD? Draw its basic structure and give its applications.
6. Write briefly about:
   (a) Channelled gate arrays

   (b) Channelless gate arrays with neat sketches.
7. Draw and explain the Antifuse Structure for programming the PAL device.
8. Explain how the I/O pad is programmed in FPGA.
9. Draw and explain the pseudo-nMOS PLA schematic for full adder and whatare the advantages and disadvantages of it.
7. Explain the gate level and function level of testing.
8. A sequential circuit with in? Inputs and ‘m’ storage devices. To test this circuit how many test vectors are required?
9. What is sequential fault grading? Explain how it is analyzed.
10. What is ATPG? Explain a method of generation of test vector.
11. Explain the terms controllability, observability and fault coverage.

12. Draw the basic structure of parallel scan and explain how it reduces the long scan chains.
13. Draw the state diagram of TAP Controller and explain how it provides the control signals for test data and instruction register.
14. (a) Explain how function of system can be tested.
    (b) Explain any one of the method of testing bridge faults.
    (c) What type of faults can be reduced by improving layout design?
15. (a) What type of defects are tested in manufacturing testing methods?
    (b) What is the Design for Autonomous Test and what is the basic device used in this?
    © What types of tests are used to check the noise margin for CMOS gates?
10. (a) What are the reasons of malfunctioning of chip? What are the different levels of testing?
    (b) Explain how a parallel scan is used for data path test.
    (c) What is mean by level sensitive of logic system?

MID & ASSIGNMENT QUESTION PAPERS

Hall Ticket No. **Question Paper Code: A1425**

CMR COLLEGE OF ENGINERRING & TECHNOLOGY

(AUTONOMOUS)

B. Tech (ECE) - SEVENTH SEMESTER - FIRST MID EXAMINATION - SEPT -2017

Subject: VLSI DESIGN **Date & Time: 07/09/2017 & 2:00 PM to 3:20 PM** **Max Marks: 25**

PART A

Answer all the following questions. (10 x 1 = 10)

1.State Moore's Law.

2.Write levels of Integrations

3.Draw the sysmbols of NMOS and PMOS Enhancement and Depletion transistors.

4.Define Figure of merit.

5.Define Synthesis

6.write different types of MOS layers.

7.draw the circuit of CMOS inverter.

8.write any three parameters of scaling.

9.Y=(AB) ',Draw the expression in pseudo CMOS logic.

10.write different types of altrnative logic gates.

PART B

Answer any 3 of the following questions. (3 x 5 = 15)

11(a)Explain the following process steps in manufacturing IC.
Oxidation and lithography
(b) Explain NMOS fabrication process.

12. Find the drain-to-source current versus voltage relationship of Ids vs Vds for a nMOS transistor.

13.a)Explain about CMOS inverter along with its regions of operation
b)What is stick diagram? Draw the stick diagram for NOR gate and NAND gate.

14.(a) What are lamda based design rules. explain?
(b) Write limitations of Scaling.

15.(a) Explain Switch Logic and transmission gate.
(b) Explain pseudo CMOS and Clocked CMOS logic.

Hall Ticket No. | | | | | | | | | | | **Question Paper Code: A1425**

CMR COLLEGE OF ENGINERRING & TECHNOLOGY

(AUTONOMOUS)

B. Tech (ECE) - SEVENTH SEMESTER – FIRST MID EXAMINATION - SEPT -2017

Subject: VLSI DESIGN **Date & Time: 07/09/2017 & 2:00 PM to 3:20 PM** **Max Marks: 25**

PART A

Answer all the following questions. (10 x 1 = 10)

1 State Moore's Law.

2.Write levels of Integrations

3.Draw the sysmbols of NMOS and PMOS Enhancement and Depletion transistors.

4.Define transconduntances.

5.Define simulation

6.write different types of MOS layers.

7.draw the circuit of NMOS inverter.

8.write any three limitations of scaling.

9.Y=(A+B)',Draw the expression in pseudo CMOS logic.

10.write different types of altrnative logic gates.

PART B

Answer any 3 of the following questions. (3 x 5 = 15)

11. Explain CMOS (nWell and Twin Tub) fabrication process.

12.Derive the expression for Zpu/Zpd ratio of one inverter driven by another inverter.

13(a) Explain BiCMOS inverter with neat sketch.

(b) Explain latch up problem in CMOS circuits.

14(a) Draw Schematic and Stick diagram for the expression Y=(AB+CD)'.

(b) Derive the some of scaling parameters.

15(a) Explain Switch Logic and pass transistor.

(b) Explain domino CMOS and NP CMOS logic.

Hall Ticket No. **Question Paper Code: A1425**

CMR COLLEGE OF ENGINERRING & TECHNOLOGY (AUTONOMOUS)

B. Tech (ECE) - SEVENTH SEMESTER - FIRST ASSIGNMENT EXAMINATION - SEPT -2017

Subject: VLSI DESIGN **Max Marks: 05**

Answer all the following questions. (5 x 1 =5)

1. Explain the fabrication process of nMOS.
2. Derive the expression for Zpu/Zpd ratio of one inverter driven by another inverter using pass transistor.
3. (a) Draw the stick diagram and layout for the logic expression Y=((A+B)CD)'.

   (b) Explain BiCMOS inverter in detail.
4. Derive all the parameters and limitations of scaling of CMOS.
5. Explain all alternative logic gates in detail.