GURUNANAK INSTITUTE OF TECHNOLOGY

157/F, Nilgunj Road, Panihati
Kolkata -700114
Website: www.gnit.ac.in
Email: info.gnit@jisgroup.org

Approved by A.I.C.T.E., New Delhi
Affiliated to MAKAUT, West Bengal

OCW

VLSI & Microelectronics

Course Level: Undergraduate

Credit: 3

Prepared by:

Ms. Swagata Bhattacharya
Assistant Professor (ECE)

Dr. Sunipa Roy
Associate Professor (ECE)

Course Name: VLSI & Microelectronics
Paper Code: EC702
Total Contact Hours: 3L+1T /Week,
Total Contact Hours : 45
Credit: 4

Course Objective: Objective of the course VLSI & Microelectronics, Code : EC702 is to motivate students to design VLSI circuits in the area of digital , analog and also to encourage for the design of IC with low power and high speed .

Course Name	COs	CO Statement
VLSI & Microelectronics(EC702)	EC702.CO1	Able to describe scale of integration – SSI ,MSI,LSI,VLSI, Moor's Law , scaling , short channel effect ,VLSI design flow, FPGA architecture and construct gate level circuit with PAL & PLA concept.
	EC702.CO2	Able to analyze CMOS inverter voltage transfer characteristics with the parameters – V_{IL} , V_{IH} , V_{OL} , V_{OH} , V_{th} and based on the knowledge of digital circuit design methodology like – CMOS , Pass transistor , TG , DCVSL , dynamic logic , NORA , able to construct schematic of combinational , sequential circuit , SRAM , DRAM cell using MOSFET
	EC702.CO3	Based on the fundamental concept of MOSFET characteristics and model , able to calculate value of resistance of current source ,MOS diode , current of current mirror circuit , voltage of references (voltage divider , threshold voltage and band gap), emulate resistance of switch capacitor circuit , gain of switch capacitor integrator and 1st order switch capacitor filter .
	EC702.CO4	With the help of MOS transistor model, able to calculate the value of parameters to design CMOS differential amplifier and two stage OP-AMP .
	EC702.CO5	Able to describe fabrication steps of IC and construct stick diagram & layout of CMOS inverter and basic gates based on lambda and micron design rules.
	EC702.CO6	Able to calculate gate delay, dynamic power, short circuit power and leakage power and total power consumption across CMOS inverter circuit based on the derived expression of delay and power.

Prerequisite:

Concept of courses Solid State Devices, Analog Electronic Circuit , Digital Electronic and Circuit.

Module –1 : Introduction to IC

1.1 Integrated Circuits

1.1.1 Advantages of Integrated Circuits

The major advantages of integrated circuits over those made by interconnecting discrete components are as follows :

1. Extremely small size – Thousands times smaller than discrete circuits. It is because of fabrication of various circuit elements in a single chip of semiconductor material.
2. Very small weight owing to miniaturised circuit.
3. Very low cost because of simultaneous production of hundreds of similar circuits on a small semiconductor wafer. Owing to mass production of an IC costs as much as an individual transistor.
4. More reliable because of elimination of soldered joints and need for fewer interconnections.
5. Lower power consumption because of their smaller size.
6. Easy replacement as it is more economical to replace them than to repair them.
7. Increased operating speed because of absence of parasitic capacitance effect.
8. Close matching of components and temperature coefficients because of bulk production in batches.
9. Improved functional performance as more complex circuits can be fabricated for achieving better characteristics.
10. Greater ability of operating at extreme temperatures.
11. Suitable for small signal operation because of no chance of stray electrical pickup as various components of an INC are located very close to each other on a silicon wafer.
12. No component project above the chip surface in an INC as all the components are formed within the chip.

1.1.2 Disadvantages of Integrated Circuits

The major disadvantages of integrated circuits over those made by interconnecting discrete components are as follows :

1) In an IC the various components are part of a small semiconductor chip and the individual component or components cannot be removed or replaced, therefore, if any component in an IC fails, the whole IC has to be replaced by a new one.

2) Limited power rating as it is not possible to manufacture high power (say greater than 10 W) ICs.

3) Need of connecting inductors and transformers exterior to the semiconductor chip as it is not possible to fabricate inductor and transformers on the semiconductor chip surface.

4) Operation at low voltage as ICs function at fairly low voltage.Quite delicate in handling as these cannot withstand rough handling or excessive heat.

5) Need of connecting capacitor exterior to the semiconductor chip as it is neither convenient nor economical to fabricate capacitances exceeding 30pF. Therefore, for higher values of capacitance, discrete components exterior to IC chip are connected.

6) High grade P-N-P assembly is not possible.

7) Low temperature coefficient is difficult to be achieved.

8) Large value of saturation resistance of transistors.

9) Voltage dependence of resistor and capacitors.

10) The diffusion processes and other related procedures used in the fabrication process are not good enough to permit a precise control of the parameter values for the circuit elements. However, control of the ratios is at a sufficiently acceptable level.

1.1.3 The Limitations of an Integrated Circuit:

1)A single integrated circuit can only work when it is connected to the corresponding peripheral components and is provided power source.

2) There are many transistors but few inductors, resistors and capacitors in integrated circuits, because making those inductors need to use large areas of silicon which result in high cost.

3) Once the integrated circuit is manufactured, the internal circuit couldn’t be changed, unlike the discrete component circuit. Thus, the whole integrated circuit can only be replaced when one of the components in the integrated circuit is damaged.

4) The integrated circuit can't be used alone, which need to be combined with discrete components and form a practical circuit.

No technological advancement ever comes without a downside. Integrated circuits have limitations that engineers must consider when designing an electronic device or system.While some components are easy to fabricate onto chips, other components defy the IC manufacturing process. Inductors, except for components with extremely low values (in the nanohenry range), constitute a prime example. Devices

using ICs must generally be designed to work with discrete inductors (coils) external to the ICs themselves. This constraint need not pose a problem, however. Resistance-capacitance (RC) circuits can do

1.2 Scale of Integration – SSI , MSI ,LSI,VLSI ,ULSI

IC design has evolved from single transistors to SSI (small-scale integration), to MSI (medium-scale integration), to LSI (large-scale integration) and to VLSI (Very Large Scale Integration). An IC is normally classified by either by the number of transistors it has, such as LSI, VLSI, and so on, or by the size of the transistor (covered in Chapter 4). Typical pitch sizes are 1, 1.5 and 2 mm (2 micros). Table 1.1 outlines the typical applications for the different classifications. Table 1.1 Design classifications. Type No. of transistors Typical applications SSI 1-100 Logic gates, op-amps, linear applications. MSI 100-1 000 Registers, filters, and so on. LSI 1 000-100 000 8-bit microprocessors, up to 64 kbit ROMs and RAMs, Analogue-to-Digital converters, and so on VLSI 100 000-500 000 16/32-bit microprocessors, up to 256 kbit ROMs/RAMs, signal processors. ULSI† >500 000 64-bit microprocessors, 8 Mbit RAMs, real-time and image processors.GSI* >10 000 000 64 Mbit RAMs, integrated multi-processors. † ULSI represents ultra-large scale integration * GSI represents gigantic scale integration.

first IC was invented around 1959 by Jack Kilby.

There after integrity has come like SSI,MSI,LSI and VLSI

In SSI(Small Scale Integration) —10–100 transistors/chip or 3 - 30 gates /chip(logic gates,flip flops)

In MSI(Medium Scale Integration) —100–1000 transistors/chip or 30 - 300 gates /chip(counters,multiplexers,registers)

In LSI(Large Scale Integration) —1000–10,000 transistors/chip or 300 - 3000 gates /chip(8 bit processors)

In VLSI(Very Large Scale Integration) —10,000–1,00,000 transistors/chip or morethan 3000 gates /chip.(16 bit and 32 bit processors)

In ULSI(Ultra Large Scale Integration) —10power 6 –10 power 7 transistors/chip(smart sensors,VR reality modules)

1.3 Moore's Law

Moore's Law asserts that the number of transistors on a microchip doubles every two years, though the cost of computers is halved. In other words, we can expect that the speed and capability of our computers will increase every couple of years; and we will pay less for them. Another tenet of Moore's Law is that this growth in the microprocessor industry is Exponential meaning that it will expand steadily and rapidly over time.Understanding Moore's Law

In 1965, Gordon E. Moore—the co-founder of Intel (NASDAQ: INTC)—postulated in a magazine article that the number of transistors that can be packed into a given unit of space will double about every two years. (Now, however, doubling of installed transistors on silicon chips occurs closer to every 18 months instead of every two years.) Gordon Moore did not call his observation "Moore's Law," nor did he set out to create a "law." Moore made that statement based on noticing emerging trends in chip manufacturing at Intel. Moore's insight became a prediction, which in turn became the golden rule known as Moore's Law.

Moore's Law proved to be true. For decades following Gordon Moore's original observation, Moore's Law has guided the semiconductor industry in long-term planning and setting targets for research and development (R&D). Moore's Law has been a driving force of technological and social change, productivity, and economic growth that are hallmarks of the late-twentieth and early twenty-first centuries. Moore's Law—Nearly 60 Years, Still Strong. More than 50 years later, we feel the lasting impact and benefits of Moore's Law in many ways .Moore's Law implies that computers, machines that run on computers, and computing power all become smaller and faster with time, as transistors on integrated circuits become more efficient. Chips and transistors are microscopic structures that contain carbon and silicon molecules, which are aligned perfectly to move electricity along the circuit faster. The faster a microchip processes electrical signals, the more efficient a computer becomes. Costs of these higher-powered computers eventually decrease by about 30% per year because of lower labor costs.Practically every facet of a high-tech society benefits from Moore's Law in action. Mobile devices, such as smartphones and computer tablets would not work without tiny processors; neither would video games, spreadsheets, accurate weather forecasts, and global positioning systems (GPS).Moreover, smaller and faster computers improve transportation, health care, education, and energy production—to name but a few of the industries that have progressed because of the power of computer chips. **[Important: Moore's Law may reach its natural end in the 2020s.]**Experts agree that computers should reach the physical limits of Moore's Law at some point in the 2020s. The high temperatures of transistors eventually would make it impossible to create smaller circuits. This is because cooling down the transistors takes more energy than the amount of energy that already passes through the transistors. In a 2005 interview, Moore himself admitted that his law "can't continue forever. It is the nature of exponential functions," he said, "they eventually hit a wall. "Shrinking transistors have powered advances in computing for more than half a century, but soon engineers and scientists must find other ways to make computers more capable. Instead of physical processes, applications and software may help improve the speed and efficiency of computers. Cloud computing, wireless communication, the Internet of Things, and quantum physics all may play a role in the future of computer tech innovation.The vision of an endlessly empowered and interconnected future brings both challenges and benefits. Privacy and security threats are growing concerns. In the long run, however, the advantages of ever-smarter computing technology ultimately can help keep us healthier, safer, and productive.Examples of Moore's Law abound everywhere we turn today. For instance, you likely have experienced the need to purchase a new computer or phone more often than you thought—say every two-to-four years—either because it was too slow, would not run a new application well, or for other

reasons. This is a phenomenon of Moore's Law that we all know well.Perhaps, however, Moore's Law—or its impending death—is most painfully present at the chip manufacturers themselves; as these companies are saddled, not only with making our computing chips but building them with increasing capacity against the physical odds. Even Intel is competing with itself and its industry to create what ultimately may not be possible.In 2012, with its 22-nanometer (nm) processor, Intel was able to boast having the world's smallest and most advanced transistors in a mass-produced product. In 2014, Intel launched an even smaller, more powerful 14nm chip; and currently, the company is struggling to bring its 10nm chip to market.For perspective, one nanometer is one-billionth of a meter, smaller than the wavelength of visible light. The diameter of an atom ranges from about 0.1 to 0.5 nanometers.

1.4 Scaling of MOSFET

The reduction of the dimensions of a MOSFET has been dramatic during the last three decades. Starting at a minimum feature length of 10 mm in 1970 the gate length was gradually reduced to 0.15 mm minimum feature size in 2000, resulting in a 13% reduction per year. Proper scaling of MOSFET however requires not only a size reduction of the gate length and width but also requires a reduction of all other dimensions including the gate/source and gate/drain alignment, the oxide thickness and the depletion layer widths. Scaling of the depletion layer widths also implies scaling of the substrate doping density.In short, we will study simplified guidelines for shrinking device dimensions to increase transistor density & operating frequency and reduction in power dissipation & gate delays.

Two types of scaling are common:

1) constant field scaling and

2) constant voltage scaling.

1.4.2 Constant field scaling and constant voltage scaling

Constant field scaling yields the largest reduction in the power-delay product of a single transistor. However, it requires a reduction in the power supply voltage as one decreases the minimum feature size. Constant voltage scaling does not have this problem and is therefore the preferred scaling method since it provides voltage compatibility with older circuit technologies. The disadvantage of constant voltage scaling is that the electric field increases as the minimum feature length is reduced. This leads to velocity saturation, mobility degradation, increased leakage currents and lower breakdown voltages. After scaling, the different Mosfet parameters will be converted as given by table below: Before Scaling After Constant Field Scaling After Constant Voltage Scaling. Dennardat al. presented their pioneering research work on the scaling of MOSFET devices at the International Electron Device Meeting (IEDM) 1972 [2.14] and published a comprehensive paper on the scaling of MOS transistors in 1974 [2.15], from which the "constant field scaling" theory has emerged. The basic principle which they employ is that in order to increase the performance of a MOSFET we must reduce linearly the size of the transistor,

together with the supply voltage, and increase the doping concentration in a way which keeps the electric field in the device constant - hence the name “constant field scaling” (figure 2.5).Constant field scaling yields the largest reduction in the power-delay product of a single transistor. However, it requires a reduction in the power supply voltage as one decreases the minimum feature size. Constant voltage scaling does not have this problem and is therefore the preferred scaling method since it provides voltage compatibility with older circuit technologies. The disadvantage of constant voltage scaling is that the electric field increases as the minimum feature length is reduced. This leads to velocity saturation, mobility degradation, increased leakage currents and lower breakdown voltages.

1.4.2 Short Channel Effect

So far our discussion was based upon the assumptions that channel was long and wide enough, so that “edge” effects along the four sides was negligible, longitudinal field was negligible and electric field at every point was perpendicular to the surface. So we could perform one-dimensional analysis using gradual channel approximation. But in devices where channel is short longitudinal field will not be negligible compared to perpendicular field. So in that case one-dimensional analysis gives wrong results and we will have to perform dimensional analysis taking into account both longitudinal and vertical fields. (which is out of the scope this course)

A channel called a short channel when

(i) When junction (source/drain) length is of the order of channel length.

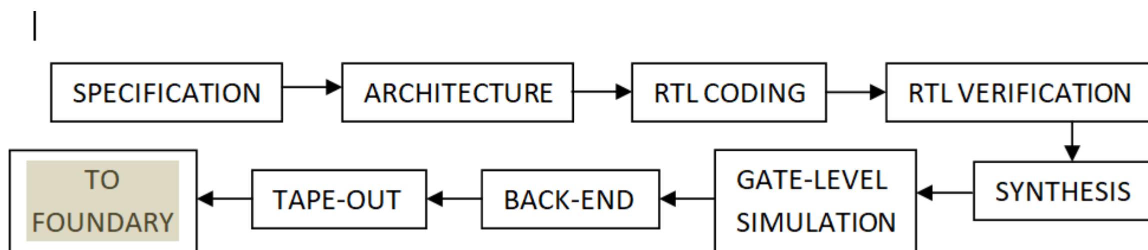
(ii) L is not much larger then the sum of the drain and source depletion width.

We have shown below the comparative graphs of **I-V** characteristics for both long channel and short channel length MOSFETs. From graph, it can be clearly concluded that when the channel becomes short, the current in saturation region becomes linearly dependent on applied drain voltage rather than being square dependent.

1.5.1 VLSI Design Flow

The chip design includes different types of processing steps to finish the entire flow. For each and every step, the design process requires a dedicated EDA tool. These tools have the flexibility to import or export different types of files.

The picture below shows the various steps of the design flow:



Description of each Step

SPECIFICATION: This is the crucial step as it will decide the future of the product. Lots of activity goes on to gather the market requirement. One may take feedback from potential customers on what they are looking for or what the expectations are. Once this done, the specification sheet along with finer technical details are shared to next team.

ARCHITECTURE: Now this is the step where main work starts. With the help of specification, design engineers decide the architecture and a layout is created for the IC using EDA tools.

RTL CODING: RTL is an acronym for register transfer level. This is where the detailed system specifications is converted into VHDL or Verilog language (Hardware Description Language) showing how data is transferred from register to register. This also goes through functional verification process in next step.

RTL VERIFICATION: Register Transfer level is one of the important step which ensures that the design is logically correct without major timing errors. It is very advantageous to perform this step at early stage. A testbench file may be used to verify the design using EDA tools.

SYNTHESIS: In this process RTL is transferred to netlist (gate level netlist). This is done with the help of FPGA/CPLD/ASIC hardware tools. These target boards may be accessed using IDE's provided by different vendors.

GATE-LEVEL SIMULATION: The verification of gate level simulation of the logic generated is very important. In this step various kinds of checks are included like: functionality check, timing check and physical analysis check.

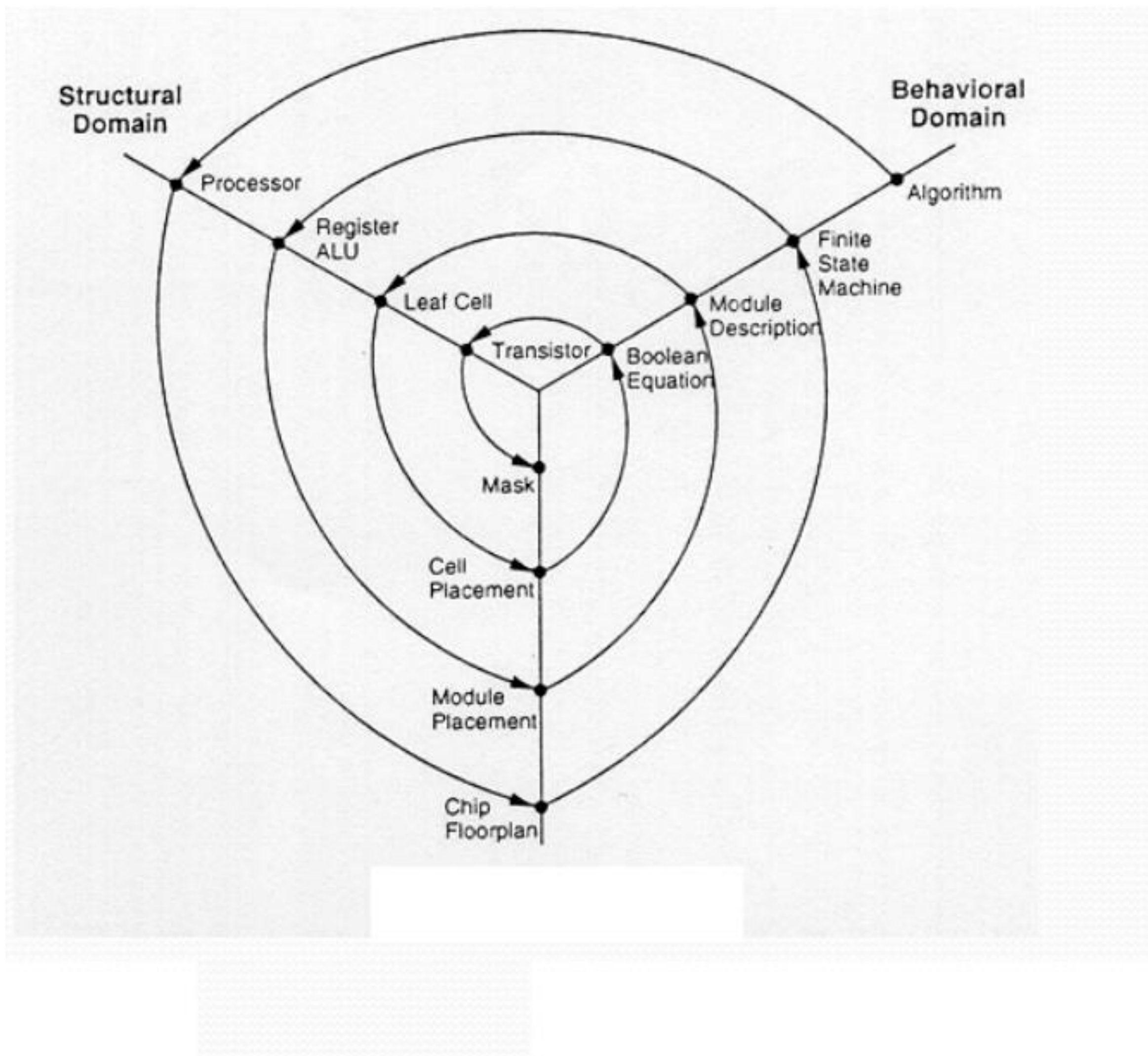
BACK-END: Here the final design after synthesis is given to the IC manufacturer.

TAPE-OUT: It is the process under back-end only where the final result of FRONT-END is provided to the manufacturer in form of photomask. Then the manufacturer performs wafer processing, packaging, testing, and delivery of samples to test the physical IC.

TO FOUNDRY: once the samples are tested and verified, then the design is sent for mass production.

1.5.2 Y CHART

(Y-chart representation)



The three different domains are behavioural, structural and physical/geometry domain which are on radial axis. Each of the corresponding domains can be further divided into levels of abstraction using concentric rings and each of the domain falling within the circle forms a group and keep going on in a top down fashion towards the centre of the core.

Let's discuss each domain in a bit detail:-

1] Physical/geometry domain:-At the physical domain it is first necessary to subdivide a large system into small ASIC sized pieces which is referred as physical partition. After that the small blocks are arranged together in the form of clusters. Then the next step

is floor plan where it is necessary to find out the approximate location of each module or block in the chip. After that we can move on to module layout and cell layout where each unit of block to be placed on chip and the connections between the cells and the block is defined.

2] Behavioural domain:-In this domain the behaviour of the entity can be modelled using procedural code which can represent the circuit at a higher level of abstraction. In this domain the first hierarchy is system where the behaviour of the architecture is defined. Then a set of algorithm needs to be defined for processing the different parameters and a set of flowchart. The third step is RTL where the coding needs to be done and then the logic level where we are concerned with 0 and 1 level, After which the transfer function is defined. Generally the people who are more interested how the system works entered into this field.

3] Structural domain:-In this domain a particular block is connected across with set of signals or netlist and here in this case we are more interested in the structure. Generally the people in this field are more interested to know how a system works. So all CPU design, verification related works come in this field as this domain is more inclined towards processors and memories. So the above description summarises the concept behind Y chart. Now let's consider the Y chart first without any kind of circle drawn on it. The point which I am trying to make is that when we draw the circle, each one of them cuts through different points across the three different domains forming a group and the process is repeated as we keep drawing the circle.

1.6 IC Classification

1.6.1 Standard IC and ASIC

Design of electronic appliances gives you a wide range of design options. Each option has its pros and cons. You will often realise that the design is not limited by electronic parameters. It is more likely limited by power, type of sensors, communication bandwidth, space, and form factors you have at your disposal and – last but not least – your product's market sales price and thereby your product manufacturing cost.

The optimal solution is like a Gordian knot. It is a compromise between many factors mostly driven by your access to skills and resources. Knowing the exact volume and product life time would be an ideal scenario. However, more often the basis of your decisions is mere predictions.

The **Gordian Knot** is a legend of Phrygian Gordium associated with Alexander the Great and a metaphor for a problem that is not easily solved.
Many projects start at demonstrator level by use of standard components (off-the-shelf Integrated Circuits – IC). When the demonstrator or the first series is accepted, the cost must be cut down in order to meet market expectations. Other projects start with a product idea that requires high integration due to the form factor (e.g. portability) or available supply of power. These are the typical scenarios where it is relevant to evaluate whether a custom made integrated circuit (also called Application Specific Integrated Circuit or ASIC) is the right choice.

An ASIC is an integrated circuit (IC) made for one specific purpose – a custom made chip so to speak. This is quite the opposite to a standard electronic component where you buy off-the-shelf ICs and only use part of the functions on the chip.

Most integrated circuits consist of a "die", a lead frame or substrate, and a package to protect the circuit. A die is semiconductor material with a functional electronic circuit on board – see Picture 1.

An ASIC is designed and manufactured as any other IC. However, it is 100% focused on the specific application it is made for. This means that the ASIC is optimised for performance and thus eliminates the redundant functions, area and cost that is unavoidable when using a standard component.

Standard components

A standard IC is general-purpose electronics i.e. designed for a broad range of use. The IC is designed in exactly the same way as an ASIC, by use of the same technologies, libraries and design tools.

The standard component is excellent for fast prototyping and to obtain a short time to market. The main drawback is that when you ramp up in volume, the only advantage you get is a small price reduction. If you need several standard components to do the job, you will have a lot of devices to keep track on.

Another drawback is that you do not know for how long this standard component is on the market. Once a standard component becomes obsolete, you need to redesign your product. This is often critical for a company as it steels precious resources from your other important design tasks and creates a vulnerable gap in the product supply.

An ASIC gives you full control of the supply chain and you avoid nasty surprises raised by obsolete components.

Comparison	***Standard component***	***ASIC***	***Comment***

Time to market	++++	÷÷	*Standard components are ideal for prototyping and small series.*
Non-Recurring Expenses (NRE)	+	÷÷÷	*Introduction cost is low for standard component applications.*
Unit cost	÷÷÷	+++	*With the optimal set-up for a given volume the unit cost for ASIC is very low especially due to no redundant functions and overhead*
Scalability	÷÷÷	+++	*With an ASIC it is easy to increase the volume without extra effort.*
Power consumption	÷÷	++++	*The design for an optimal power consumption is the real strength of ASIC technology with no redundant functions and dedicated technology for the application.*
Form factor	÷÷÷	++++	*ASIC is ideal for portable and low power applications due to its small form factor and power consumption.*
Control of supply chain	÷÷÷	+++	*Transparency in the supply chain:*

			Access to wafer fab, assembly and test house is provided with the ASIC supply chain. For a standard component the distributor is the only source.

Table 1: Comparison between a standard component versus an ASIC

Best performance and lowest unit cost & power consumption

The choice is simple when full custom design, large volume and small form factor are required. Large volume and smallest form factor equal lowest price for ASIC.

Optimised design enables best performance, the lowest power consumption and the smallest form factor. However, there are other situations where an ASIC is relevant. We will in the following address various relevant parameters.

Form factor

With integration it is obvious that the dimension of the electronics shrinks. What is often not so obvious is that the packaging of the electronics can be optimised for an even better form such as low component height which is standard in mobile electronics such as phones and tablets.

Ensure your production

A product which consist of several standard components and has a long product lifetime will inevitably need a redesign when one or more standard components become obsolete.

A reduced Bill of Material (BoM) will prevent many challenges of redesign. Thus could an ASIC be the solution as an ASIC is an integration of many components into one single chip.

The challenge can be to control the entire ASIC supply chain. This is, however, one of DELTA's core competences which we have handled for many years for many customers world-wide.

Protect your know-how

Integration of the electronic design into an ASIC first of all provides you with the smallest form factor. In addition, an ASIC also secure copy protection significantly because reverse engineering will be very difficult.

Cost parameters for integrated circuits

Most electronic components consist of a die, a lead frame or substrate, and packaging. More advanced components can have multiple dies and/or components in the same package.

Test and logistics (i.e. handling, shipment, quality assurance and control) are required to provide functioning components. For a typical IC the cost structure looks like this:

Die	Technology;Die Size
Packaging	*Type of package*
Testing	*Test time, yield*
Handling	*Logistics, shipment, QA/QC*

Table 2: Cost factors for ICs

The above cost structure is the same for a standard component and an ASIC. However, the cost of marketing, distribution and storage must be added to the standard components. This is the manufacturing cost structure, but what about the sales price?

Price structures for a standard component versus ASIC

The development cost (also called the Non-Recurring Engineering (NRE) investment) for an ASIC is high. The lead time is significant compared to buying an off-the-shelf standard component from a distributor.

The price of a standard component is low because many users split the NRE and together they reach a significant volume. The drawback is that distribution of components is a cost factor. In general, the gross margin for standard devices is 60%.

The price of an ASIC is defined by the specific volume that you as a customer forecasts and order. A typical gross margin is 20-40%.

When to choose an ASIC over a standard component

The obvious reply is "*When the volume is high enough*" (see Picture 2). In Picture 2 a standard component and an ASIC are compared in a graph where price versus volume is the only parameter.

However, there are many other parameters to consider such as: Do you need a small solution? Do you need low power? Do you need a long product life time? Many customers do ASIC implementation even though the volume does not meet the crossing point shown in Picture 2.

1.6.2 Programming Array Logic

Programmable Array Logic (PAL) is a commonly used programmable logic device (PLD). It has programmable AND array and fixed OR array. Because only the AND array is programmable, it is easier to use but not flexible as compared to Programmable Logic Array (PLA). PAL’s only limitation is number of AND gates.

PAL consist of small programmable read only memory (PROM) and additional output logic used to implement a particular desired logic function with limited components.

Comparison with other Programmable Logic Devices:

Main difference between PLA, PAL and ROM is their basic structure. In PLA, programmable AND gate is followed by programmable OR gate. In PAL, programmable AND gate is followed by fixed OR gate. In ROM, fixed AND gate array is followed by programmable OR gate array.

Describing the PAL structure (programmable AND gate followed by fixed OR gate).

Example: Realize the given function by using PAL:

Any form from sum of product (SOP) form or product of sum (POS) can be used for realization of a boolean function.

There are three inputs A, B, C and three functions X, Y, Z. Using sum of product (SOP) terms to express the given function as follows:-

X(A, B, C)=\sum(2, 3, 5, 7)

Y(A, B, C)=\sum(0, 1, 5)

Z(A, B, C)=\sum(0, 2, 3, 5)

Following Truth table will be helpful in understanding function on number of inputs:

A	B	C	X	Y	Z
0	0	0	0	1	1
0	0	1	0	1	0
0	1	0	1	0	1
0	1	1	1	0	1
1	0	0	0	0	0
1	0	1	1	1	1
1	1	0	0	0	0
1	1	1	1	0	0

Finding X, Y, Z:

Look for high min terms (function value is equal to 1 in case of SOP) in each function output:

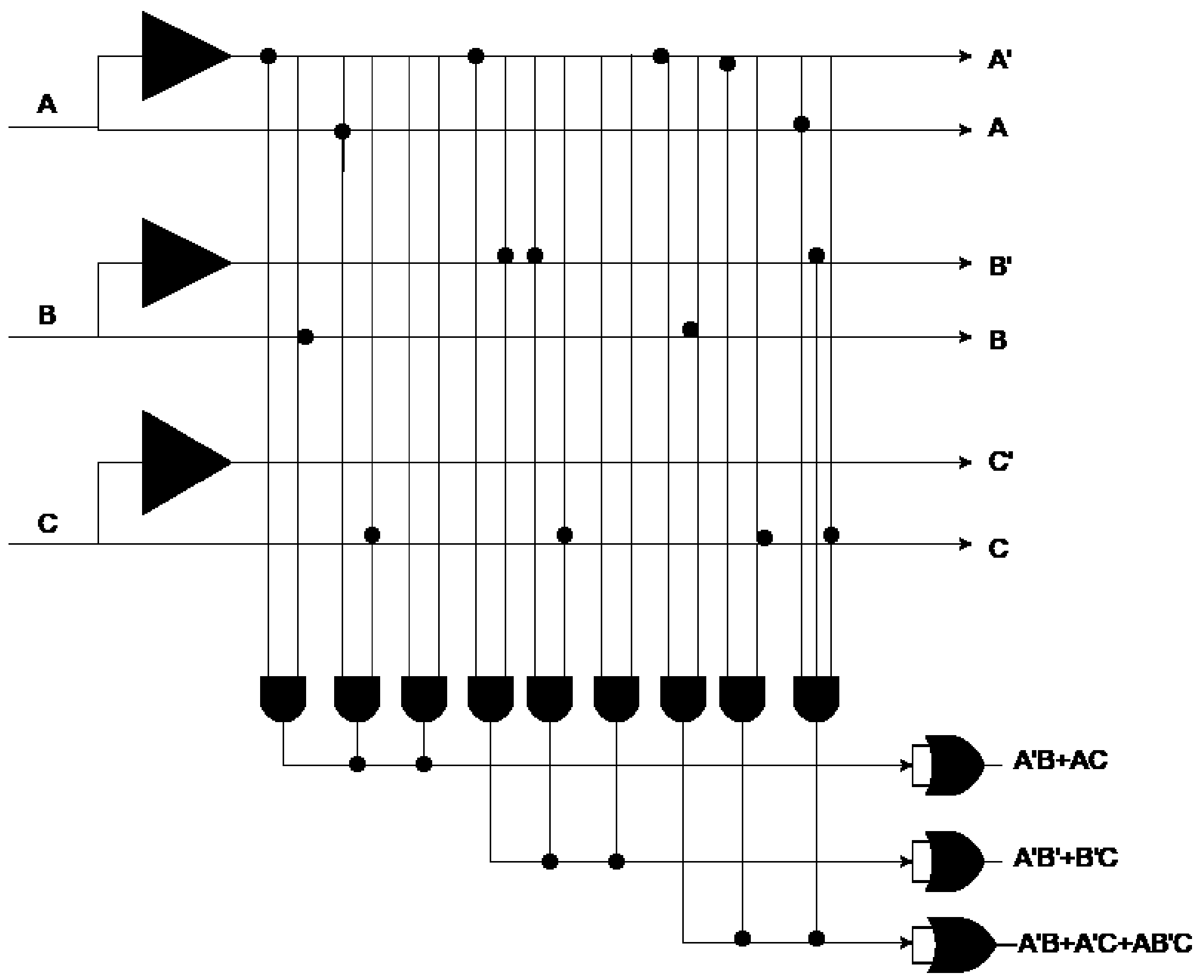
X = A’B + AC

Y = A’B + B’C

Z = A’B + A’C + AB’C

AND array has been programmed but have to work with fixed OR array as per requirement. Desired lines will be connected in PLDs.

Advantages of PAL: Highly efficient,Low production cost as compared to PLA,Highly secure,High Reliability,Low power required for working,More flexible to design.



1.6.3 PLA (**PLA (programmable logic array)**

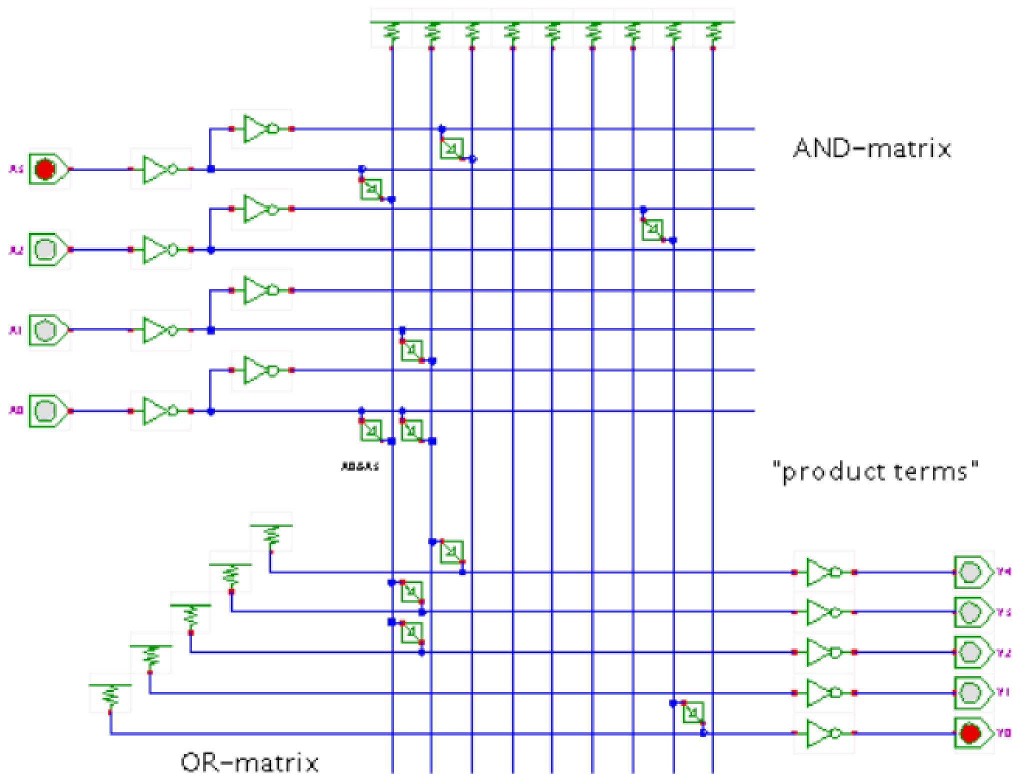


Figure 1.6.3.1: PLA

Circuit Description

Figure above shows the structure of a PLA or programmable logic array.Logically, a PLA is a circuit that allows implementing Boolean functions in sum-of-product form. The typical implementation consists of input buffers for all inputs, the programmable AND-matrix followed by the programmable OR-matrix, and output buffers.As shown in the applet, the input buffers provide both the original and the inverted values of each PLA input (here called A0, A1, A2, A3). The input lines run horizontally into the AND matrix, while the so-called product-term lines run vertically. Therefore, the size of the AND matrix is twice the number of inputs times the number of product-terms. A weak pullup-resitors ensures that the default value of all product-term signals is high.Each place in the AND-matrix holds a small diode. Depending on the programming data, this diode is left unconnected, or connected to its input-line and product-line. While the unconnected diode will do nothing, the product-term line will be driven low by the connected diode whenever the corresponding input-line is low. This is the wired-AND operation: a product term will only remain high when none of the (connected!) input-lines is driven low. For example, the leftmost product-line (vertical) in the applet is connected via diodes to both A0 and

inverted A3. Therefore, P0 = (!A3 & A0). A similar argument shows that the second product-line P1 = (A2 & A0).The same structure is repeated in the output (OR-) matrix: the output-lines are driven by weak-pullup resistors, but can be driven low by product-term lines, whenever the corresponding diode is connected during programming.

For example, the upper output line Y4 is connected to both the first (leftmost) and second product-term lines, so that Y4 = P0 | P1, or Y4 = (!A3&A0) | (A2&A0). Similarly, Y0 = P6 | P7 | P8 = A2 | A1 | A0. The remaining product terms and outputs are unused, so that Y3 = Y2 = Y1 = 0.The main advantage of the PLA structure is that a very compact and space-efficient realization is possible in NMOS technology. Small self-conducting (enhancement-mode) NMOS transistors are used for the pullup-resistors, while a depletion-mode NMOS transistor is placed at each location in the AND- and OR-matrices. The first-level metal mask decides whether to connect the transistors or not. The total-size of a PLA (excluding buffers) is calculated from twice the number of input lines plus the number of output terms times the number of product terms. For many functions, PLAs are much more compact that the discrete realization based on traditional gates. However, the pullup-transistors imply that a PLA draws a (rather large) static current. As low-power consumption is a primary concern in many current devices, PLAs are not as popular in current (CMOS-) technology integrated circuits as they were in the early era of VLSI.

1.6.4 FPGA Architecture

1.6.4.1 Introduction

The full form of FPGA is "Field Programmable Gate Array". It contains ten thousand to more than a million logic gates with programmable interconnection. Programmable interconnections are available for users or designers to perform given functions easily. A typical model FPGA chip is shown in the given figure. There are I/O blocks, which are designed and numbered according to function. For each module of logic level composition, there are CLB's (Configurable Logic Blocks).

CLB performs the logic operation given to the module. The inter connection between CLB and I/O blocks are made with the help of horizontal routing channels, vertical routing channels and PSM (Programmable Multiplexers).

The number of CLB it contains only decides the complexity of FPGA. The functionality of CLB's and PSM are designed by VHDL or any other hardware descriptive language. After programming, CLB and PSM are placed on chip and connected with each other with routing channels.

The general FPGA architecture consists of three types of modules. They are I/O blocks or Pads, Switch Matrix/ Interconnection Wires and Configurable logic blocks (CLB). The basic FPGA architecture has two dimensional arrays of logic blocks with a means for a user to arrange the interconnection between the logic blocks. The functions of an FPGA architecture module are discussed below:

- CLB (Configurable Logic Block) includes digital logic, inputs, outputs. It implements the user logic.
- Interconnects provide direction between the logic blocks to implement the user logic.
- Depending on the logic, switch matrix provides switching between interconnects.
- I/O Pads used for the outside world to communicate with different applications.

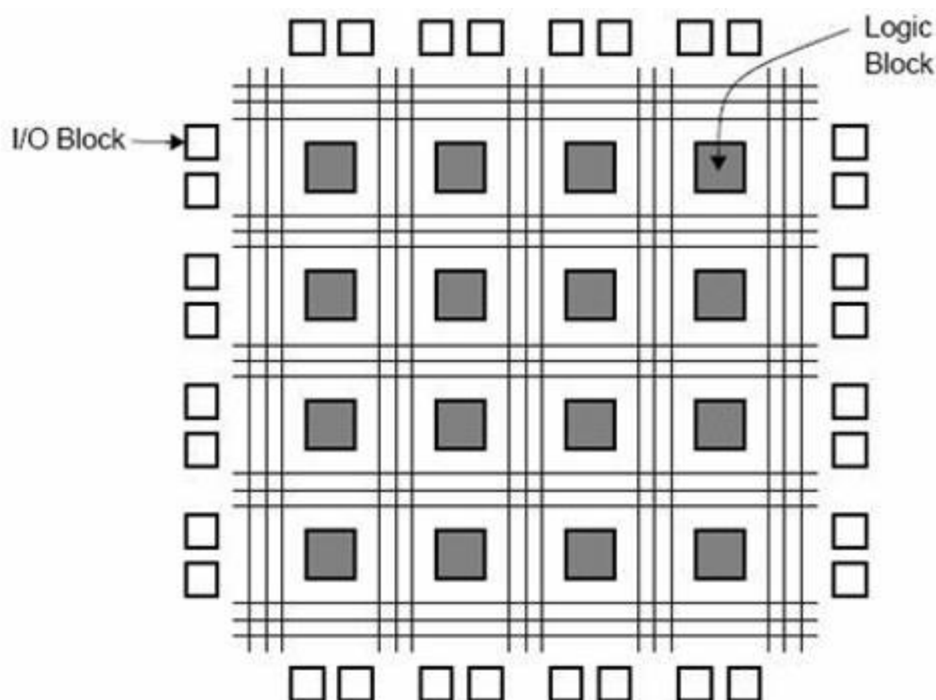


Figure 1.6.4.1:FPGA

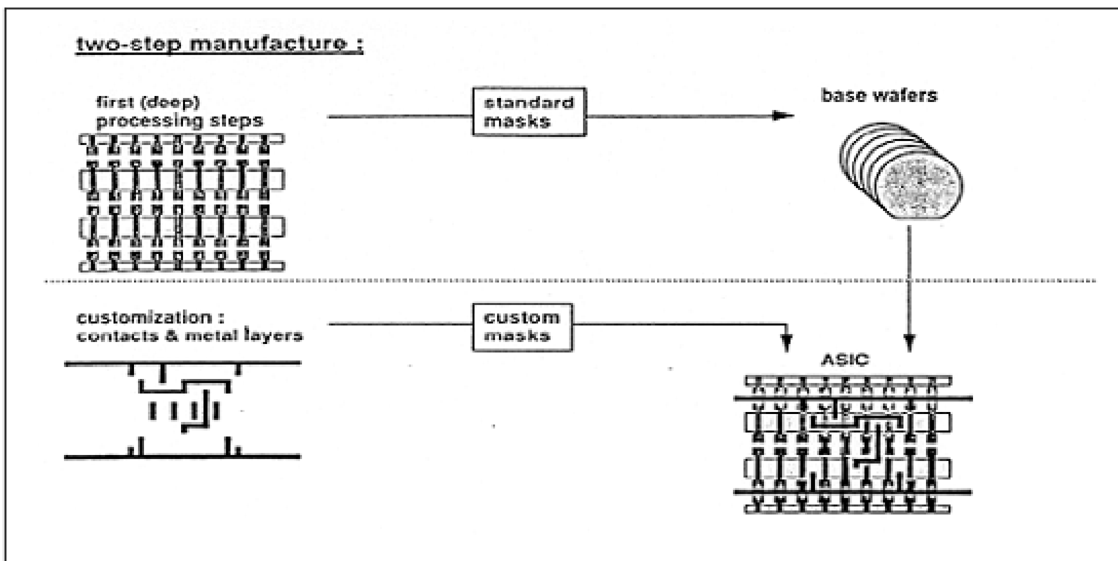
Advantages

- It requires very small time; starting from design process to functional chip.
- No physical manufacturing steps are involved in it.
- The only disadvantage is, it is costly than other styles.

1.6.4 .2 Gate Array Design

The gate array (GA) ranks second after the FPGA, in terms of fast prototyping capability. While user programming is important to the design implementation of the FPGA chip, metal mask design and processing is used for GA. Gate array implementation requires a two-step manufacturing process.

The first phase results in an array of uncommitted transistors on each GA chip. These uncommitted chips can be stored for later customization, which is completed by defining the metal interconnects between the transistors of the array. The patterning of metallic interconnects is done at the end of the chip fabrication process, so that the turn-around time can still be short, a few days to a few weeks. The figure given below shows the basic processing steps for gate array implementation.



Typical gate array platforms use dedicated areas called channels, for inter-cell routing between rows or columns of MOS transistors. They simplify the interconnections. Interconnection patterns that perform basic logic gates are stored in a library, which can then be used to customize rows of uncommitted transistors according to the netlist.

In most of the modern GAs, multiple metal layers are used for channel routing. With the use of multiple interconnected layers, the routing can be achieved over the active cell areas; so that the routing channels can be removed as in Sea-of-Gates (SOG) chips. Here, the entire chip surface is covered with uncommitted nMOS and pMOS transistors. The neighbouring transistors can be customized using a metal mask to form basic logic gates.

For inter cell routing, some of the uncommitted transistors must be sacrificed. This design style results in more flexibility for interconnections and usually in a higher density. GA chip utilization factor is measured by the used chip area divided by the total chip area. It is higher than that of the FPGA and so is the chip speed.

1.6.4 .3 **Programming Technologies**

FPGA's can be considered to be building bricks which allow desired customization of the hardware. These are a special form of PLDs with higher densities and with increased capability of implementing functionality in a shorter time span using CAD. The FPGA's are available in various flavours based on the programming technology used. These may be programmed using:

1. **Antifuse Technology,** which can be programmed only once. Devices manufactured by QuickLogic are examples of this type. Configuration is done by burning a set of fuses. These act as replacements for Application Specific ICs (ASIC) and used in places where protection of intellectual property is top priority.

2. **Flash Technology** based Programming, like devices from Actel. The FPGA may be reprogrammed several thousand times, taking a few minutes in the field itself for reprogramming and has non-volatile memory.

3. **SRAM Technology** based FPGAs, the currently dominating technology offering unlimited reprogramming and very fast reconfiguration and even partial reconfiguration during operation itself with little additional circuitry. Most companies like Altera, Actel, Atmel and Xilinx manufacture such devices.

1.6.4 .4 Configurable Logic Block

The basic building block of a Configurable Logic Block is the logic cell. A logic cell may consist of an input function generator, carry logic and a storage element. The function generators are implemented as Look Up Tables depending on the input. For example, a Xilinx Spartan II has 4 inputs LUT where each LUT can provide a 16X1 bit Synchronous RAM which can be further multiplexed using multiplexers. An LUT may also be used as a Shift register which is used to capture burst-mode data. The storage elements may be used as edge sensitive flip-flops or level sensitive latches. The arithmetic logic includes an XOR gate for full adder operation along with dedicated carry logic lines. The figure below shows an FPGA slice:

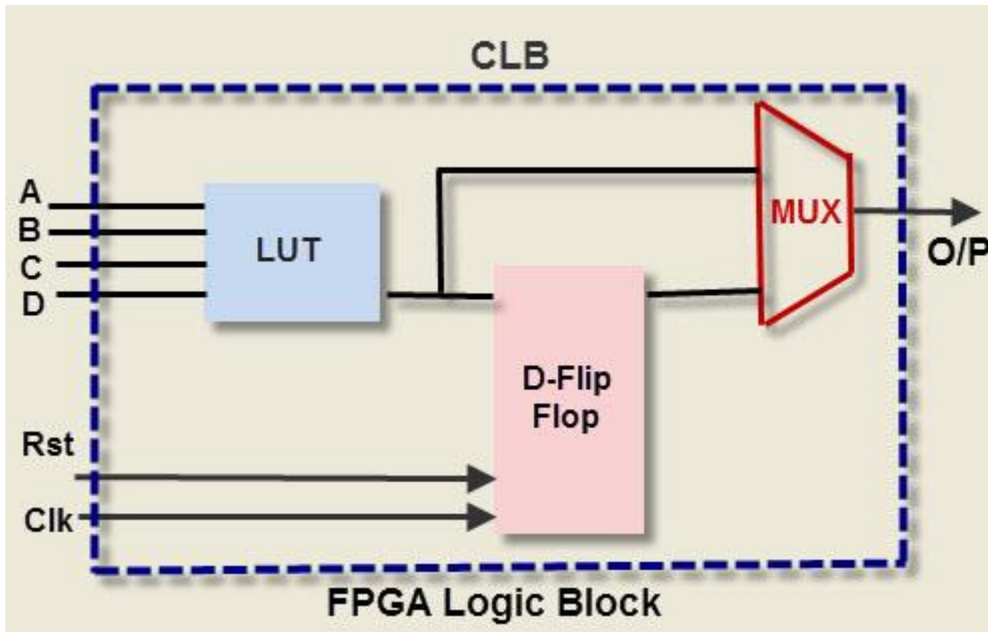


Figure 1.6.4.2 *: Figure Showing FPGA Slice*

1.6.4.5 I/O Block & Routing Matrix

Input/ Output Block: This block features inputs and outputs supporting a wide range of signaling standards and interfaces. A basic Input/ Output block is shown below:

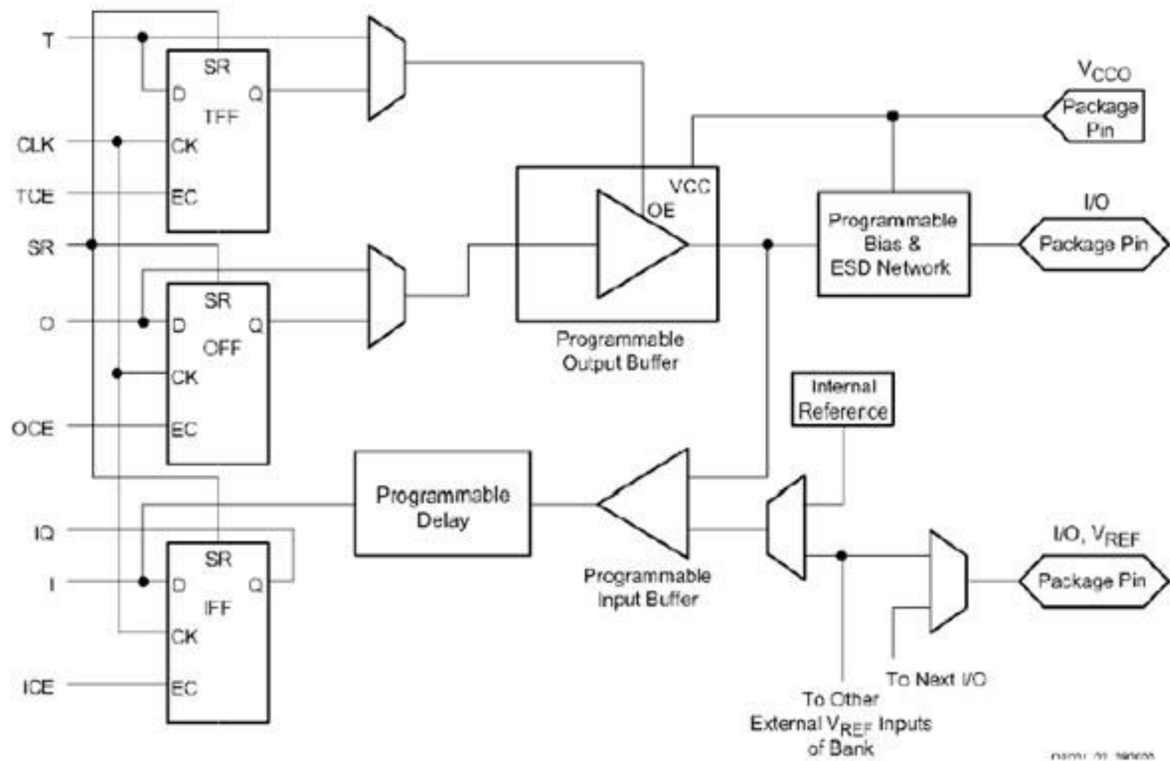


Fig. : Logic Gate Diagram Of Basic Input Output And Routing Matrix In FPGA

The buffers in the Input and output paths route the input and output signals to the internal logic and the output pads either directly or via a flip-flop. The buffer can be set to conform to various supported signaling standards which might even be user defined and externally set.

1.6.4.6 Routing Matrix

In any assembly line it is often the slowest segment which sets the overall production rate. Much in the same way, it is the route that takes the longest delay that eventually determines the performance of the entire electronic system. Thus routing algorithms are brought into place for the design of the most efficient paths to deliver optimum performance. Routing is on various levels like Local, between LUTs, flip-flops and the General Routing Matrix, General Purpose Routing between various CLBs, I/O Routing between I/O Blocks and CLBs, Dedicated Routing for a certain classes of signals for maximizing performance and Global Routing for distributing clocks and other signals with very high fanout.

1.6.4.7 Clock Distribution

High speed, low skew clock distribution is provided in most FPGAs using Primary Global Routing resources. With each clock input buffer, there is a digital Delay-Locked Loop which eliminates skew between clock input pad and internal clock input pins by adjusting the delay element, and also provides control of multiple clock domains.

FPGA families now also have large block RAM structures to complement the distributed RAM LUTs, size varying for different FPGA devices.

The design of an FPGA follows mostly the same approach as any VLSI system the chief steps being Design Entry, Behavioral Simulation, Synthesis, Post Synthesis Simulation, translation, mapping and routing, and further analysis like Timing simulation and Static Timing Analysis. On a computer, the design looks all ordered and tiled, yet imperfect placement and routing happens and leads to performance drops.

In order to increase the performance of FPGAs, more transistors could always be used. The area overhead involved in FPGAs is higher than ASIC and could possibly do with more density now that 28nm processes are also being implemented on them. Putting more transistors also means that larger designs would be possible. Leakage is a major issue with FPGAs and has been an area of interest. Use of asynchronous FPGA architecture has also shown better results coupled with pipelining technology which reduces global inputs and improves
throughput.

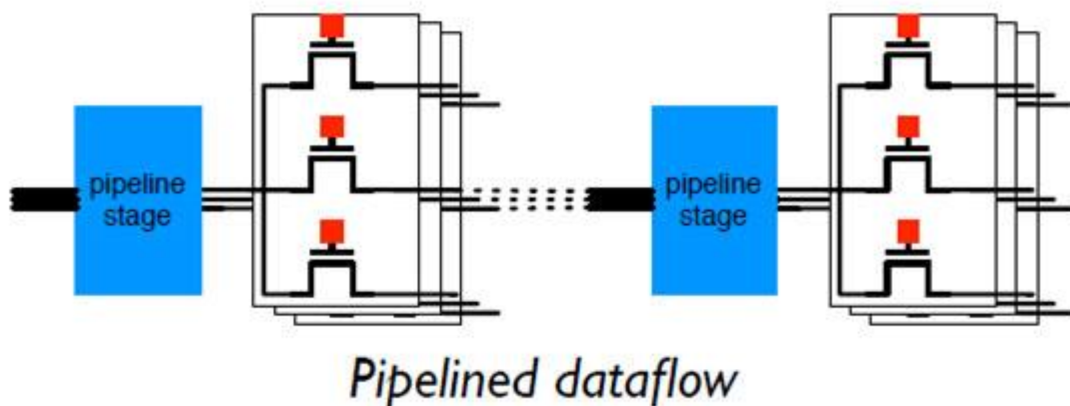


Fig. 1.6.4.3*: An Image Showing Pipeline Dataflow in FPGA*

Security used to be a major concern as the code needed to be revealed every time it was loaded on to the FPGA's, thus making FPGA's flexibility a potential threat to malicious modifications during fabrication. But, bitstream encryption has come to the rescue of FPGAs.

Often the inexperienced designers and users face this dilemma that how much powerful FPGA would be suitable for their application. Manufacturers often specify metrics like 'Gate count'. For Example, Xilinx uses 3 metrics to measure capacity of FPGA, Maximum Logic Gates, Maximum Memory Bits and Typical Gate Range. As long as these quoted metrics are consistent, migration between families is somewhat simplified, but it rarely offers subtle comparison between different vendors because of the difference of architectures and as a result of which, performance varies. A better metric is to compare the type and number of logic resources provided. In addition to it, the designer must be fully aware of what is exactly required of the device as vendors might be boasting of features which would be of least importance to the job. For example, Altera's Stratix II EP2S180 features about 1,86,576 4-Input LUTs while Xilinx Virtex-4 XC4VLX200 contains 1,78,176. However if the design needs only 177K LUTs, the latter would suffice. If RAM is the desired metric for the designer, neither the Xilinx XC4VLX200's 6Mbits nor the Altera's EP2S180's 9Mbits would be favoured over the lesser advertised, older model of XC4VFX140 standing at 9.9Mbits. So it requires thorough understanding on the part of the user who finally needs it and not what the newest product on the shelf offers.

1.6.4 .8 **Importance of FPGA**

FPGA hold promise of delivering even in harsh conditions. The cyclone devices from Altera work well in temperature ranges of -40 degrees to 85 degrees. Another factor that promotes their long term use is the long term availability. ASIC manufacturers do not agree on availability of 5 or at maximum 10 years, where as FPGAs have nearly unlimited availability even if device migrates to next generation.

These find use in microprocessor systems like the PowerPC405 embedded cores, in Digital Signal Processing as embedded multipliers and in I/O Processing like Digitally controlled Impedance. It is always better to be sure of the design and its performance by testing it on FPGA's before going in for ASIC circuits. These are employed in Defense systems and medical imaging. The possibility of evolvable hardware was revealed while implementing speech recognition on an FPGA using genetic algorithm. FPGAs being parallel processing devices find use in applications like brute force attacks used in breaking cryptographic algorithms, in convolution and FFT computations.

Sample Questions (Module -1)

1.1 Multiple Choice Questions

1.1.1) SSI stands for

a) Small scale Iterations

b) small scale Integration

c) Small size integration

d) all false

1.1.2) In MSI transistor count is

a) 1- 10

b) 30-90

c) more than 30,000

d) 100-1000

1.1.3) Moore's Law is associated with

a) Transistor count in a microchip

b) semiconductor doping

c) Impurity profile

d) all true

1.1.4) VLSI stands for

a) Very Low Scale Integration

b) Very large scale integration

c) Vast linear scale integration

d) all false

1.1.5) In behavioural domain of Y Chart

a) related with netlist formation

b) related with structural modelling

c) modelling of entity nature

d)related with cell placement

1.1.6) For a 3 input PLA ,number of AND planes are

A)3

b)1

c) 6

d) 4

1.1.7) OR plane of PLA is reponsible for

a) Creating distinct product terms

b) ANDing product terms

c) ORing product terms

d) all true

1.1.8) CLB can be observer in

a) PAL

b) PLA

c) FPGA

d) ASIC

1.1.9) Switch matrix is an important parts of

a) PAL

b) PLD

c) FPGA

d) ASIC

1.1.10) In FPGA LUT is a part of

a) I/O

b) Switch matrix

c) CLB

d) all false

1.2 Short answer type questions

1.2.1) What is an ASIC? How ASIC is different than FPGA

1.2.2) Briefly state the advantages and disadvantages of integrated circuits

1.2.3) Write a short note on Moore's law

1.2.4) State the need for MOSFET scaling.State the differences between constant field scaling and constant voltage scaling.

1.2.5) What is Y chart?State briefly about the different domains of Y Chart

1.3 Long answer type questions

1.3.1) What are the different components of FPGA?How CLB implements the logic function in FPGA? Describe the work function of switch matrix and I/O block . (2+6+3+5)

1.3.2) Implement Y1=ABC+d;Y2=AB'+C;Y3=B+C;Y4=ABCD using PLA and describe the working. (15)

1.3.3) With the help of suitable diagram explain VLSI design flow and describe each step in brief. (15)

1.3.4) What is PAL? Implement X(A, B, C)=sum(2, 3, 1, 7) ,Y(A, B, C)=sum(0, 2, 3) ,Z(A, B, C)=sum(0, 2, 4, 5) using PAL (1+14)

1.3.5)How the following functions can be implemented with CLB?Y1=AB+C;Y2=AB';Y3=ABC;Y4=Y1+Y2;Drwa the neat diagram and state each step. (15)

Module -2 : Digital VLSI Circuit Design

2.1.1 Digital VLSI Circuit Design

The basic building block of digital VLSI circuits is inverter . In the following section dealing with inverter design .

2.1.2 **Inverter Characteristics**

The logic symbol and truth table of ideal inverter is shown in figure given below. Here A is the input and B is the inverted output represented by their node voltages. Using positive logic, the Boolean value of logic 1 is represented by V_{dd} and logic 0 is represented by 0. V_{th} is the inverter threshold voltage, which is $V_{dd}/2$, where V_{dd} is the output voltage.

The output is switched from 0 to V_{dd} when input is less than V_{th}. So, for $0<V_{in}<V_{th}$ output is equal to logic 0 input and $V_{th}<V_{in}<V_{dd}$ is equal to logic 1 input for inverter.

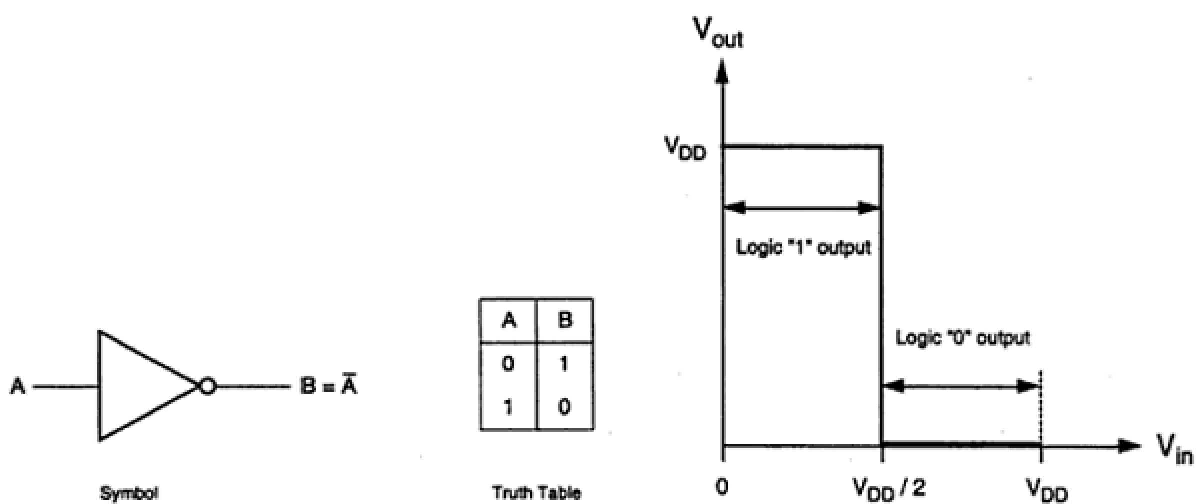


Fig 1: Ideal Inverter

The characteristics shown in the figure are ideal. The generalized circuit structure of an nMOS inverter is shown in the figure below.

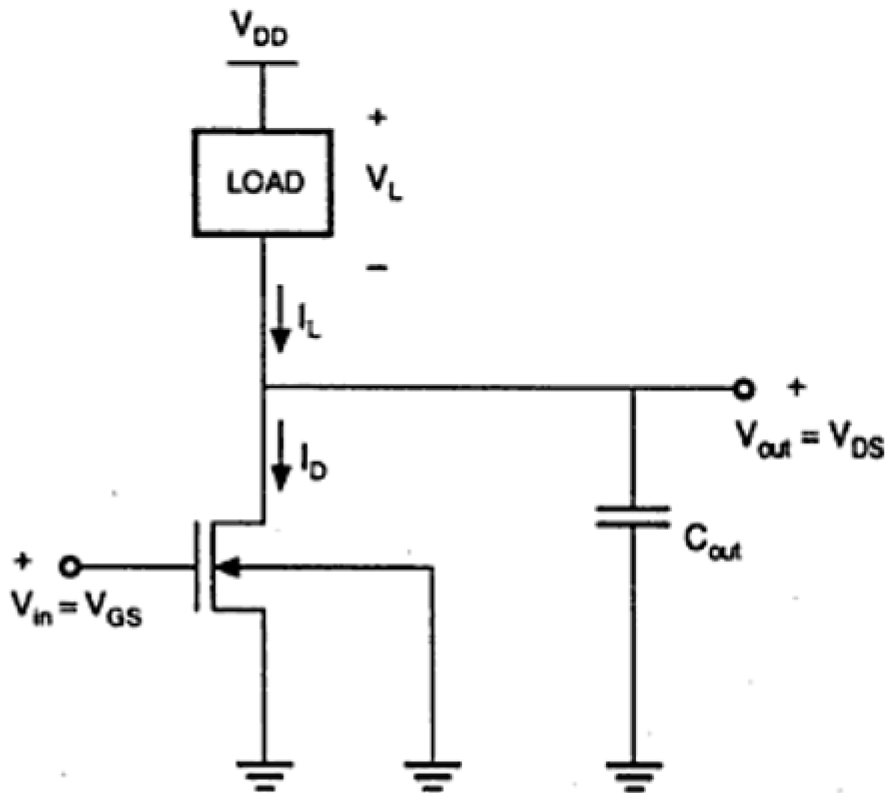


Fig 2: MOS Inverter With Load

From the given figure, we can see that the input voltage of inverter is equal to the gate to source voltage of nMOS transistor and output voltage of inverter is equal to drain to source voltage of nMOS transistor. The source to substrate voltage of nMOS is also called driver for transistor which is grounded; so $V_{SS} = 0$. The output node is connected with a lumped capacitance used for VTC.

Resistive Load Inverter

The basic structure of a resistive load inverter is shown in the figure given below. Here, enhancement type nMOS acts as the driver transistor. The load consists of a simple linear resistor R_L. The power supply of the circuit is V_{DD} and the drain current I_D is equal to the load current I_R.

Circuit Operation

When the input of the driver transistor is less than threshold voltage V_{TH} ($V_{in} < V_{TH}$), driver transistor is in the cut – off region and does not conduct any current. So, the voltage drop across the load resistor is ZERO and output voltage is equal to the V_{DD}. Now, when the input voltage increases further, driver transistor will start conducting the non-zero current and nMOS goes in saturation region.

Mathematically,

$ID=(Kn/2)\times[V_{GS}-V_{TO}]^2$

Increasing the input voltage further, driver transistor will enter into the linear region and output of the driver transistor decreases.

$ID=(Kn/2)\times 2[V_{GS}-V_{TO}]V_{DS}-V_{DS}^2$

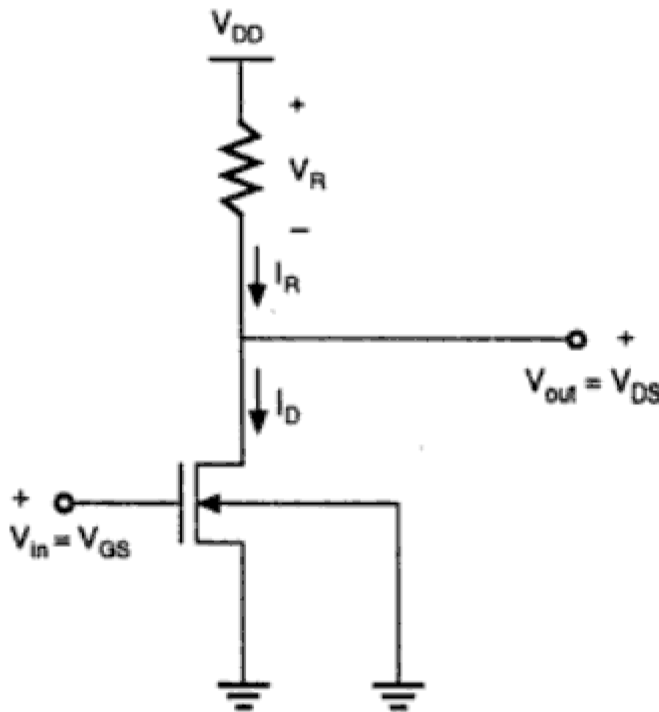


Fig 3 : Resistive Load Inverter

VTC of the resistive load inverter, shown below, indicates the operating mode of driver transistor and voltage points.

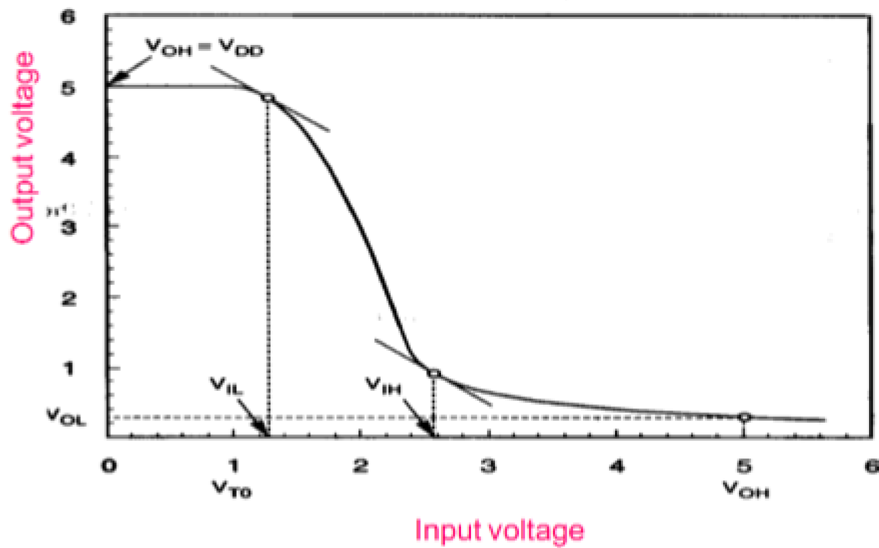


Fig 4: VTC of Resistive Load Inverter

Voltage transfer characteristic (VTC)

The VTC describing Vout as a function of Vin under DC condition .

• Very low voltage level – Vout=VOH – nMOS off, no conducting current, voltage drop across the load is very small, the output voltage is high

• As Vin increases –The driver transistor starts conducting, the output voltage starts to decrease – The critical voltage point, dVout/dVin=-1

• The input low voltage VIL

• The input high voltage VIH

• Determining the noise margins

• Further increase Vin – Output low voltage VOL, when the input voltage is equal to VOH – The inverter threshold voltage Vth

• Defined as the point where Vin=Vout

VIHmin = minimum HIGH input voltage.

VILmax = maximum LOW input voltage.

VOHmin= minimum HIGH output voltage.

VOLmax= maximum LOW output voltage.

2.1.3 CMOS Inverter – Circuit, Operation and Description

The CMOS inverter circuit is shown in the figure. Here, nMOS and pMOS transistors work as driver transistors; when one transistor is ON, other is OFF.

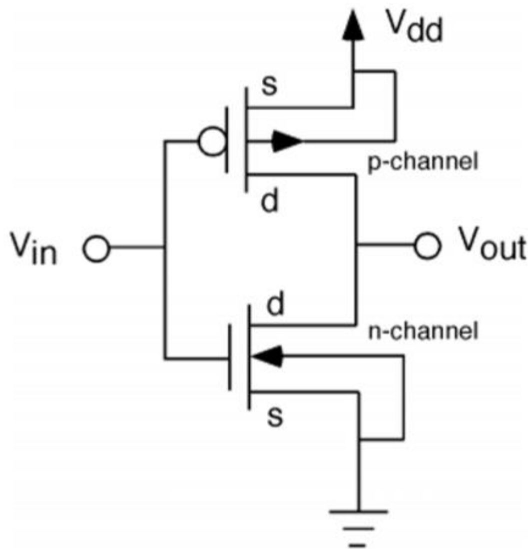


Fig 5: CMOS Inverter

This configuration is called **complementary MOS (CMOS)**. The input is connected to the gate terminal of both the transistors such that both can be driven directly with input voltages. Substrate of the nMOS is connected to the ground and substrate of the pMOS is connected to the power supply, V_{DD}.

So $V_{SB} = 0$ for both the transistors.

$$VGS,n=Vin$$

$$V_{GS},n=Vout$$

$$VDS,n=VoutVDS,n=Vout$$

And,

$$VGS,p=Vin-VDD$$

$$V_{GS,p}=V_{out}-V_{DD}$$

When the input of nMOS is smaller than the threshold voltage ($V_{in} < V_{TO,n}$), the nMOS is cut – off and pMOS is in linear region. So, the drain current of both the transistors is zero.

When the input of nMOS is smaller than the threshold voltage ($V_{in} < V_{TO,n}$), the nMOS is cut – off and pMOS is in linear region. So, the drain current of both the transistors is zero.

$$I_{D,n}=I_{D,p}=0$$

Therefore, the output voltage V_{OH} is equal to the supply voltage.

$$V_{out}=V_{OH}=V_{DD}$$

When the input voltage is greater than the $V_{DD} + V_{TO,p}$, the pMOS transistor is in the cutoff region and the nMOS is in the linear region, so the drain current of both the transistors is zero.

$$I_{D,n}=I_{D,p}=0$$

Therefore, the output voltage V_{OL} is equal to zero.

$$V_{out}=V_{OL}=0$$

The nMOS operates in the saturation region if $V_{in} > V_{TO}$ and if following conditions are satisfied.

$$V_{DS,n}\geq V_{GS,n}-V_{TO,n}$$

$$V_{out}\geq V_{in}-V_{TO,n}$$

The pMOS operates in the saturation region if $V_{in} < V_{DD} + V_{TO,p}$and if following conditions are satisfied.

$$V_{DS,p}\leq V_{GS,p}-V_{TO,p}$$

$$V_{out}\leq V_{in}-V_{TO,p}$$

For different value of input voltages, the operating regions are listed below for both transistors.

Region	V_{in}	V_{out}	nMOS	pMOS
A	$< V_{TO, n}$	V_{OH}	Cut – off	Linear
B	V_{IL}	High $\approx V_{OH}$	Saturation	Linear
C	V_{th}	V_{th}	Saturation	Saturation
D	V_{IH}	Low $\approx V_{OL}$	Linear	Saturation

The VTC of CMOS is shown in the figure below −

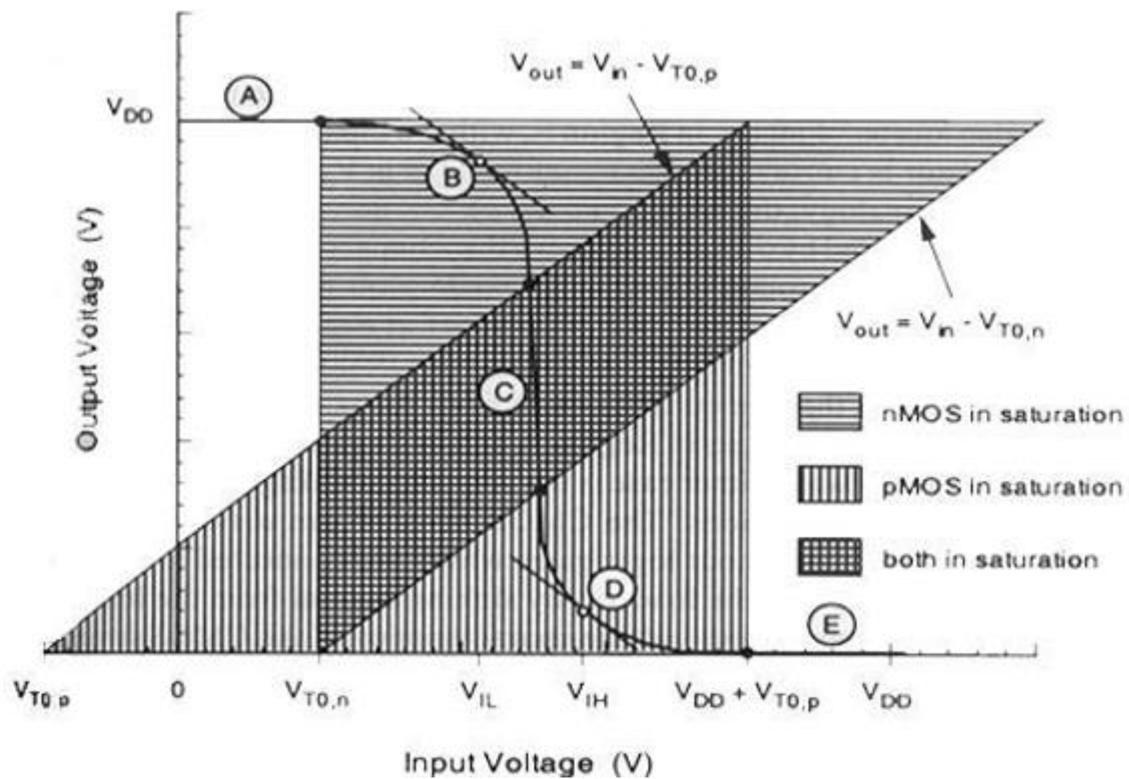


Fig 7: VTC of CMOS Inverter

Noise Margin

Noise margin is the amount of noise that a CMOS circuit could withstand without compromising the operation of circuit. Noise margin does makes sure that any signal which is logic '1' with finite noise added to it, is still recognised as logic '1' and not logic '0'. It is basically the difference between signal value and the nosie value. Refer to the diagram below.

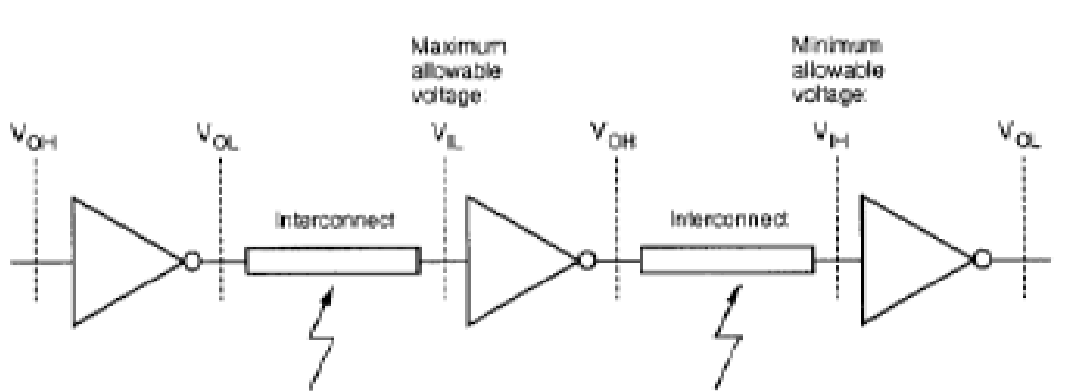


Fig 8: Chain of Inverter Depicting NOISE MARGIN

Consider the following output characteristics of a CMOS inverter. Ideally, When input voltage is logic '0', output voltage is supposed to logic '1'. Hence Vil (V input low) is '0'V and Voh (V output high) is 'Vdd'V.

Vil = 0

Voh = Vdd

Ideally, when input voltage is logic '1', output voltage is supposed to be at logic '0'. Hence, Vih (V input high) is 'Vdd', and Vol (V output low) is '0'V.

Vih = Vdd

Vol = 0

Figure below shows the N_{MH} and N_{ML} levels of two cascaded inverters. The noise margin shows the levels of noise when the gates are connected together. For the digital integrated circuits the noise margin is larger than '0' and ideally it is high.

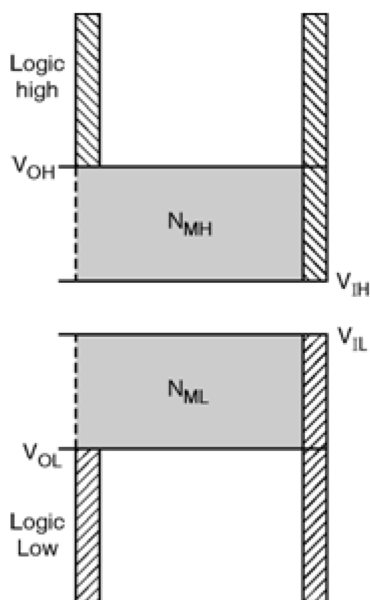


Fig 9: Noise Margin

Noise margin is a parameter closely related to the input-output voltage characteristics. This parameter allows us to determine the allowable noise voltage on the input of a gate so that the output will not be affected. The specification most commonly used to specify noise margin (or noise immunity) is in terms of two parameters- The LOW noise margin, NML, and the HIGH noised margin, NMH. With reference to Fig 4, NML is

defined as the difference in magnitude between the maximum LOW output voltage of the driving gate and the maximum input LOW voltage recognized by the driven gate. Thus, The value of NMH is difference in magnitude between the minimum HIHG output voltage of the driving gate and the minimum input HIGH voltage recognized by the receiving gate. Thus, Where,

V_{IH}min = minimum HIGH input voltage.

V_{IL}max = maximum LOW input voltage.

V_{OH}min= minimum HIGH output voltage.

V_{OL}max= maximum LOW output voltage.

Noise Margins could be defined as follows :
NMl (NOISE MARGIN low) = V_{IL} - V_{OL} = 0 - 0 = 0
NMh (NOISE MARGIN high) = V_{OH} - V_{IH} = V_{DD} - V_{DD} = 0

But due to voltage droop and ground bounce, Vih is usually slightly less than Vdd i.e. Vdd', whereas Vil is slightly higher that Vss i.e. Vss'.

Hence Noise margins for a practical circuit is defined as follows :
NMl (NOISE MARGIN low) = Vil - Vol = Vss' - 0 = Vss'
NMh (NOISE MARGIN high) = Voh - Vih = Vdd - Vdd'

Hence, if input voltage (Vin) lies somewhere between Vol and Vil, it would be detected as logic '0', and would result in an output which is acceptable. Similarly, if input voltage (Vin) lies between Vih and Voh, it would be detected as logic '1' and would result in an output which is acceptable.

2.2.1 Combinational Logic Circuit Design

Combinational logic circuits or gates, which perform Boolean operations on multiple input variables and determine the outputs as Boolean functions of the inputs, are the basic building blocks of all digital systems. We will examine simple circuit configurations such as two-input NAND and NOR gates and then expand our analysis to more general cases of multiple-input circuit structures.

Next, the CMOS logic circuits will be presented in a similar fashion. We will stress the similarities and differences between the nMOS depletion-load logic and CMOS logic circuits and point out the advantages of CMOS gates with examples. In its most general form, a combinational logic circuit, or gate, performing a Boolean function can be represented as a multiple-input, single-output system, as depicted in the figure.

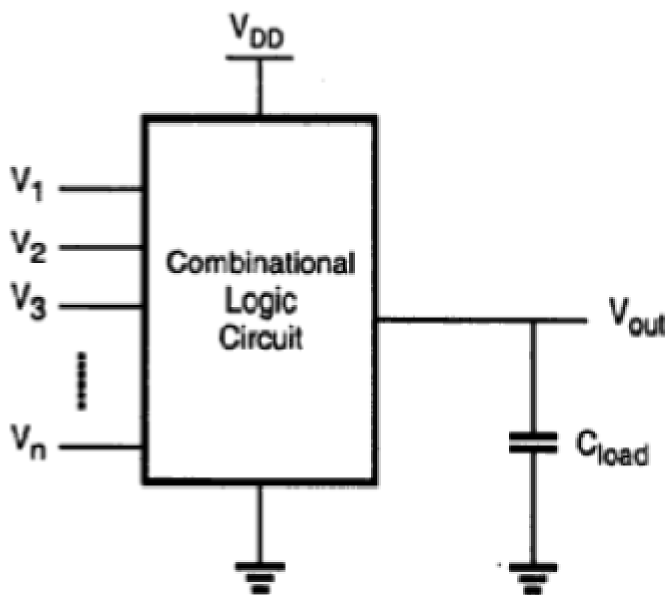


Fig 10: Combinational Logic

Node voltages, referenced to the ground potential, represent all input variables. Using positive logic convention, the Boolean (or logic) value of "1" can be represented by a high voltage of VDD, and the Boolean (or logic) value of "0" can be represented by a low voltage of 0. The output node is loaded with a capacitance C_L, which represents the combined capacitances of the parasitic device in the circuit.

2.2 Combinational Logic Circuit Design (3L+5T): Circuit design using Static CMOS style – basic gates , design of circuit for product of sum(POS) and sum of product (SOP) expression, Complex logic circuit , full adder ; Circuit design using pseudo NMOS logic , DCVSL Logic , TG Logic , Pass Transistor Logic , Complementary pass transistor logic , Dynamic logic , domino logic , NORA logic .

2.2.2 COMPLEMENTARY CMOS

Static CMOS gates are implemented by using combination of two networks, the pull up network (PUN) and pull down network (PDN).Static CMOS is characterized by very good current driving capabilities and highnoise margins. In Static CMOS design, at every point in time, each gate output is connected to either Vddor Vss via a low-resistance path. Also, the outputs of the gate assume at all times the value of the Booleanfunction implemented by the circuit.A Static CMOS gate is a combination of two networks, the pull up network (PUN) and the pull downnetwork (PDN). The function of the PDN is to provide a connection between the output and Vdd when theoutput of the logic gate is supposed to be 1. Similarly, the PDN connects the output to Vss when the outputis expected to be 0. The PUN and PDN networks are constructed in a mutually exclusive manner such thatone and only one of the networks are conducting in steady state.The Static CMOS gates have rail-to-rail swing, no static power dissipation. The speed of the static CMOS circuit depends on the transistor sizing and the various parasitic that are involved with it. The problem withthis type of implementation is that for N fan-in gate 2N number of transistors are required, i.e., more arearequired to implement logic. This has an impact on the capacitance and thus the speed of the gate.

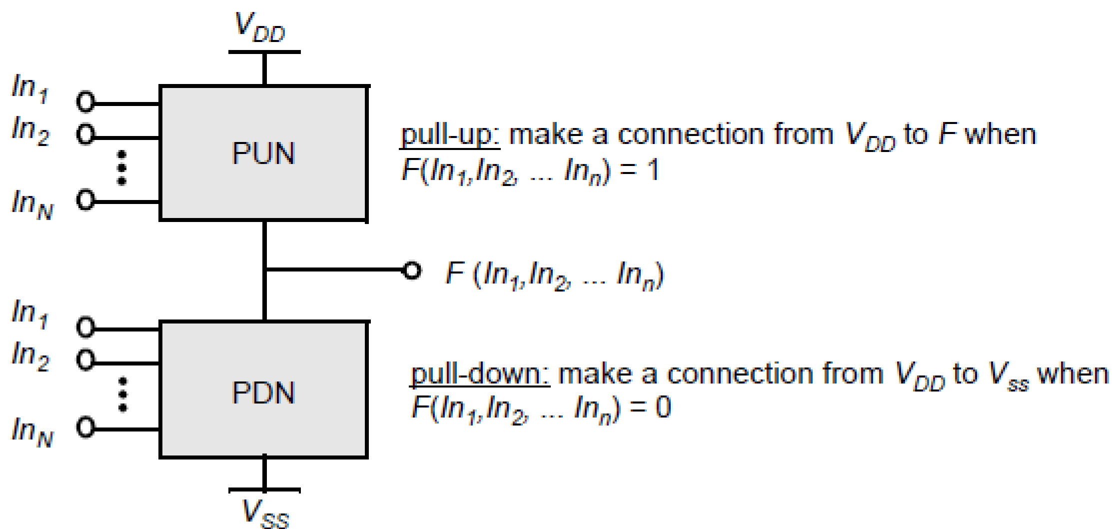


Fig 11: PUN and PDN Network

In constructing PUN and PDN following point should be considered.

1. A transistor can be assumed as a switch controlled by gate voltage. NMOS operates on positive gatevoltage and PMOS operate on application of negative gate voltage.
2. The PDN is realied by using NMOS while PUN is by using PMOS transistors. This is due to the fact that NMOS produce strong 0s and PMOS device generate strong 1s

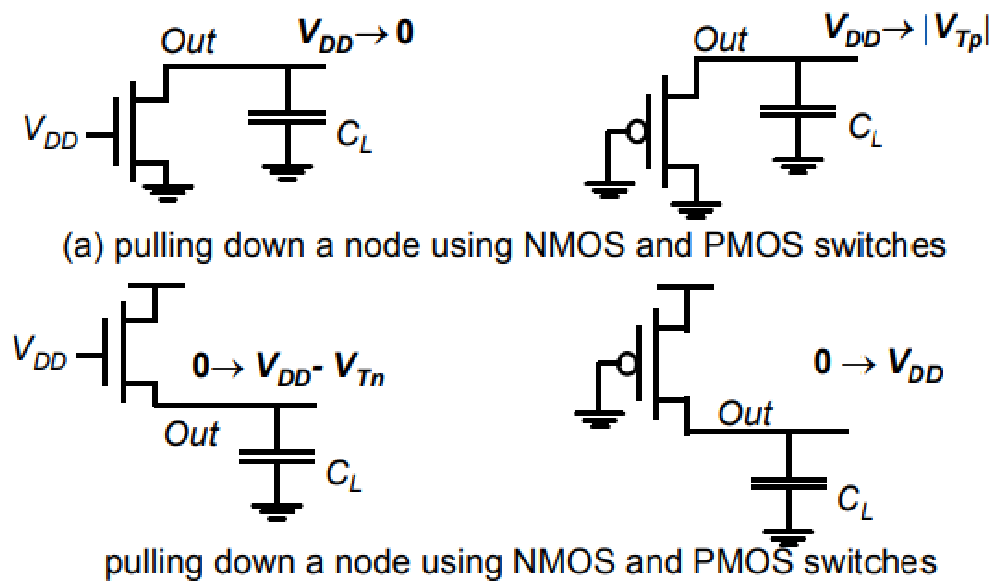


Fig 12: NMOS and PMOS Switches depicting strong '0' and strong '1'.

A set of construction rules can be derived to construct logic functions (Figure 6.4). NMOS devices connected in series corresponds to an AND function. With all the inputs high, the series combination conducts and the value at one end of the chain is transfered to the other end. Similarly, NMOS transistors connected in parallel represent an OR function. A conducting path exists between the output and input terminal if at least one of the inpurs is high. Using similar arguments, construction rules for PMOS networks can be formulated. A series connection of PMOS conducts if both inputs are low, representing a NOR function ($\bar{A}.\bar{B} = \overline{A+B}$), while PMOS transistors in parallel implement a NAND ($\bar{A}+\bar{B}=\overline{A.B}$)

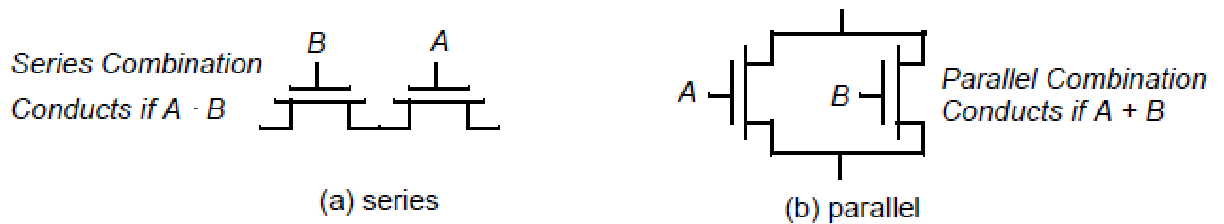


Fig 13: Series and Parallel Combination using NMOS

• Using De Morgan's theorems (($A + B$) = $A·B$ and $A·B$ = $A + B$), it can be shown that the pull-up and pull-down networks of a complementary CMOS structure are *dual* networks. This means that a parallel connection of transistors in the pull-up network corresponds to a series connection of the corresponding devices in the pull-down network, and vice versa. Therefore, to construct a CMOS gate, one of the networks (e.g., PDN) is implemented using combinations of series and parallel devices. The other network (i.e., PUN) is obtained using duality principle by walking the hierarchy, replacing series subnets with parallel subnets, and parallel subnets with series subnets. The complete CMOS gate is constructed by combining the PDN with the PUN.

• The complementary gate is naturally *inverting*, implementing only functions such as NAND, NOR, and XNOR. The realization of a non-inverting Boolean function (such as AND OR, or XOR) in a single stage is not possible, and requires the addition of an extra inverter stage.

The number of transistors required to implement an N-input logic gate is $2N$.

2.2.4 INPUT CMOS NAND GATE

TRUTH TABLE

V_A	V_B	V_{out}
Low	Low	High
Low	High	High
High	Low	High
High	High	Low

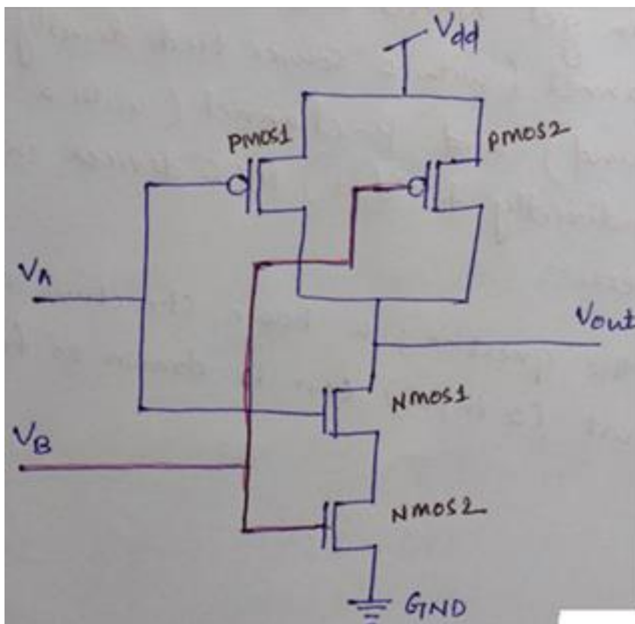


Fig 14 : 2 Input CMOS NAND

The above drawn circuit is a 2-input CMOS NAND gate. Now let's understand how this circuit will behave like a NAND gate. The circuit output should follow the same pattern as in the truth table for different input combinations.

Case-1 : V_A – Low & V_B – Low

As V_A and V_B both are low, both the pMOS will be ON and both the nMOS will be OFF. So the output V_{out} will get two paths through two ON pMOS to get connected with V_{dd}. The output will be charged to the V_{dd} level. The output line will not get any path to the GND as both the nMOS are off. So, there is no path through which the output line can discharge. The output line will maintain the voltage level at V_{dd}; so, High.

Case-2 : V_A – Low & V_B – High

V_A – Low: pMOS1 – ON; nMOS1 – OFF

V_B – High: pMOS2 – OFF; nMOS2 – ON

pMOS1 and pMOS2 are in parallel. Though pMOS2 is OFF, still the output line will get a path through pMOS1 to get connected with V_{dd}. nMOS1 and nMOS2 are in series. As nMOS1 is OFF, so V_{out} will not be able to find a path to GND to get discharged. This in turn results the V_{out} to be maintained at the level of V_{dd}; so, High.

Case-3 : V_A – High & V_B – Low

V_A – High: pMOS1 – OFF; nMOS1 – ON

V_B – Low: pMOS2 – ON; nMOS2 – OFF

The explanation is similar as case-2. V_{out} level will be High.

Case-4 : V_A – High & V_B – High

V_A – High: pMOS1 – OFF; nMOS1 – ON

V_B – High: pMOS2 – OFF; nMOS2 – ON

In this case, both the pMOS are OFF. So, V_{out} will not find any path to get connected with V_{dd}. As both the nMOS are ON, the series connected nMOS will create a path from V_{out} to GND. Since, the path to ground is established, V_{out} will be discharged; so, Low.

In all the 4 cases we have observed that V_{out} is following the exact pattern as in the truth table for the corresponding input combination.

2 Input NOR Gate

TRUTH TABLE

V_A	V_B	V_{out}
Low	Low	High
Low	High	Low
High	Low	Low
High	High	Low

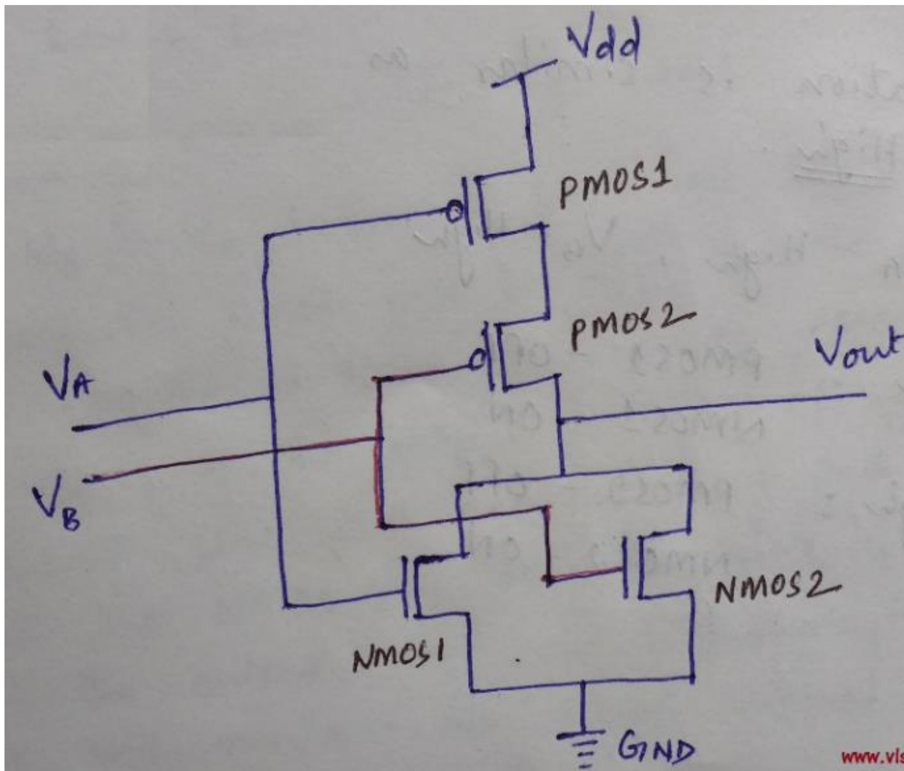


Fig 15: 2 Input CMOS NOR

The above drawn circuit is a 2-input CMOS NOR gate. Now let's understand how this circuit will behave like a NOR gate.

Case-1 : V_A – Low & V_B – Low

V_A – Low: pMOS1 – ON; nMOS1 – OFF

V_B – Low: pMOS2 – ON; nMOS2 – OFF

Path establishes from V_{dd} to V_{out} through the series connected ON pMOS transistors and V_{out} gets charged to V_{dd} level. No path from V_{out} to GND. Therefore, no discharging and hence V_{out} will be High.

Case-2 : V_A – Low & V_B – High

V_A – Low: pMOS1 – ON; nMOS1 – OFF

V_B – High: pMOS2 – OFF; nMOS2 – ON

In this case path establishes from V_{out} to GND through nMOS2, but no path to V_{dd}. So, V_{out} would get discharged and will be at level Low.

Case-3 : V_A – High & V_B – Low

V_A – High: pMOS1 – OFF; nMOS1 – ON

V_B – Low: pMOS2 – ON; nMOS2 – OFF

The explanation is similar as case-2. V_{out} will be at level Low.

Case-4 : V_A – High & V_B – High

V_A – High: pMOS1 – OFF; nMOS1 – ON

V_B – High: pMOS2 – OFF; nMOS2 – ON

No path to V_{dd}. Path establishes from V_{out} to GND. So, V_{out} will be at level Low.

In all the 4 cases we have observed that V_{out} is following the expected value as in 2 input NOR gate truth table.

For the design of 'n' input NAND or NOR gate:

Let's say n = 3

In case of NAND gate, 3 pMOS will be connected in parallel and 3 nMOS will be connected in series, and other way around in case of 3 input NOR gate. The same pattern will continue even if for more than 3 inputs.

2.2.5 COMPLEX LOGIC CIRCUIT

Complex gates can be realized at transistor level – which is advantageous as the gate delay is smaller for one complex gate than for the series connection of several simple gates realizing the same function. Usually the number of inputs is limited to 4 (the number of transistors in series between the ground and supply is limited). The realized logic function can be any combination of the AND and NOR functions and there is always an inversion at the output:

$y = \overline{(A + B)C}$

$y = \overline{AB + CD}$

$y = \overline{(A + B)CD}$

Let's design the complex gate realizing the logic function y = $\overline{(A + B)C}$ First the pull-down network (PDN) is created. The OR function is realized by two n-type FETs connected in parallel. The AND function is realized by two n-type FETs connected in series.

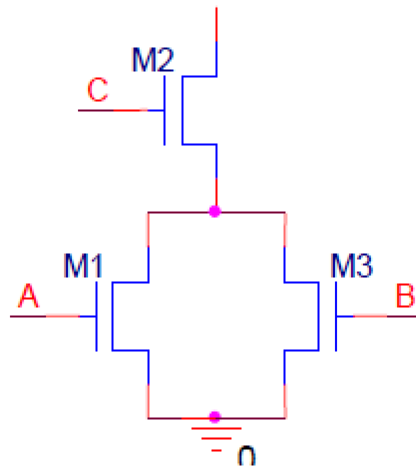


Fig 16: PDN Network

Next the pull-up network (PUN) is designed with p-type transistors. The PUN has to create a current path between the supply rail and the output for every logic 1 of the logic function. This can be done by creating the dual network of the PDN. In the dual network every series connection is turned into a parallel connection and vica versa.

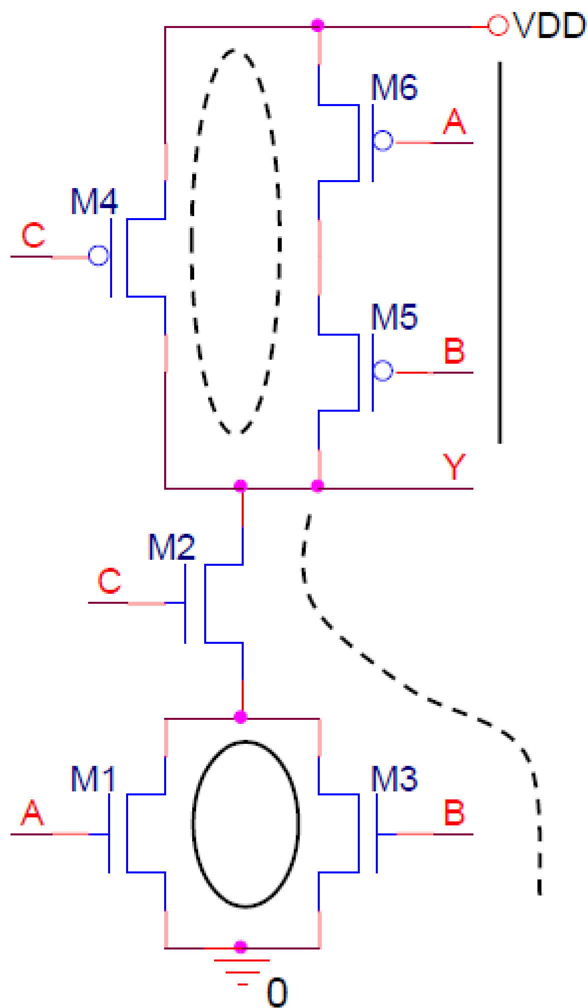


Fig 17:PUN and PDN Network Combined for Combinatinal CMOS Logic

As the p-type transistors conduct when the input is logic 0, the function has to be inverted using the De Morgan laws.

In this case:

$$y = \overline{C(A\ +\ B)} = \overline{C} + \overline{A\ +\ B} = \overline{C} + \overline{AB}$$

As it can now be seen, the two methods yield the same results.

2.3.1FULL ADDER

Adder is a circuit in order to operate for a given three one bit inputs A, B, C and two one bit outputs sum and carry.

A basic full adder cell in digital computing systems has three 1-bit inputs (A, B & C) andtwo 1- bit outputs (Sum and Carry). These outputs can be expressed in many different logicexpressions. Therefore, many full adder circuits can be designed using the differentexpressions.Table 1: Truth table of full adderThe logical Boolean expressions using truth table (table 1) between the logic inputs andlogic outputs are also expressed as:

$SUM = A\oplus B\oplus C$ ....

$SUM = \bar{C} . (A\oplus B) + C.\overline{(A\oplus B)}$.....

$COUT = A.B + C\,(A\oplus B)$.............

$= C.(A\oplus B) + A.\overline{(A\oplus B)}$................

A	B	C	Sum	Carry
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

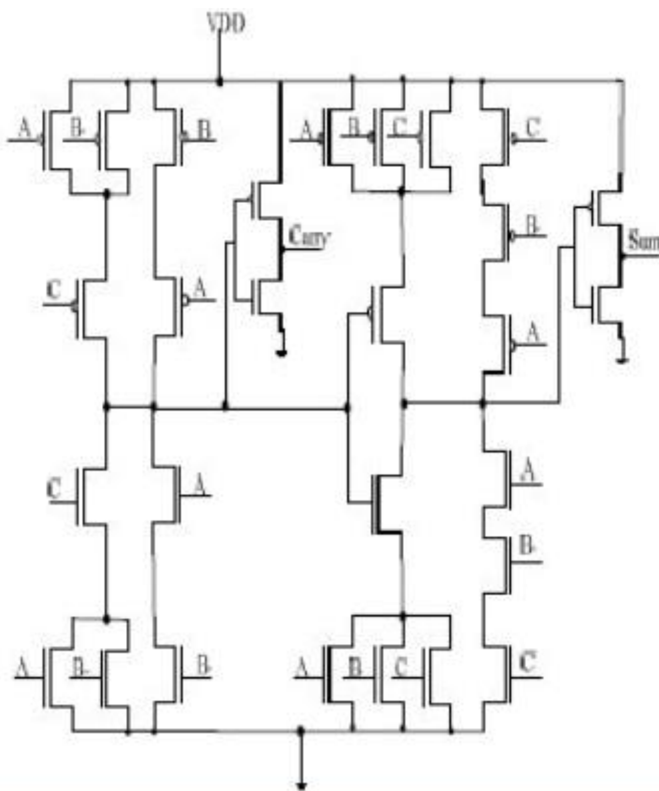


Fig 18: CMOS Full Adder

2.3.2Pseudo NMOS

Using a PMOS transistor simply as a pull-up device for an n-block as shown in Fig. below is called pseudo-NMOS logic. Note, that this type of logic is no longer ratio-less, i.e., the transistor widths must be chosen properly, i.e., The pull-up transistor must be chosen wide enough to conduct a multiple of the n-block's leakage and narrow enough so that the n-block can still pull down the output safely:

$$I_{off,n}F_{in} < W_p I_{on,p} < I_{on,n}/F_{in} \tag{A.21}$$

The advantage of pseudo-NMOS logic are its high speed (especially, in large-fan-in NOR gates) and low transistor count. On the negative side is the static power consumption of the pull-up transistor as well as the reduced output voltage swing and gain, which makes the gate more susceptible to noise. At a second glance, when pseudo-NMOS logic is combined with static CMOS in time critical signal paths only, the overall speed improvement can be substantial at the cost of only a slight increase of static-power consumption. Furthermore, when the gate of the pull-up transistor is connected to a appropriate control signal it can be turned off, i.e., pseudo-NMOS supports a power-down mode at no extra cost.

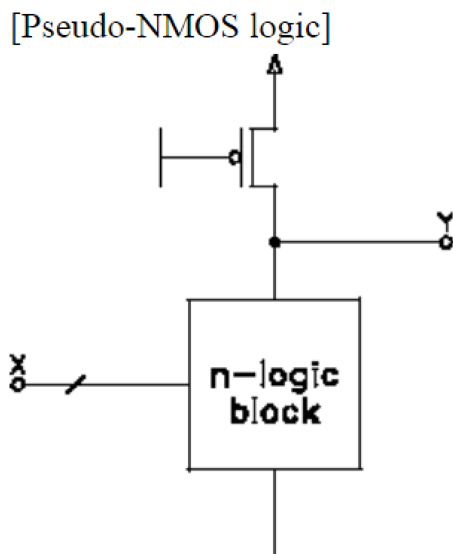


Fig 19: Peseudo NMOS

The name ``pseudo-NMOS" originates from the circumstance that in the older NMOS technologies a depletion mode NMOS transistor with its gate connected to source was used as a pull-up device.

2.3.3 DCVSL Logic

The DCVSL circuit is illustrated in Figure below 20.

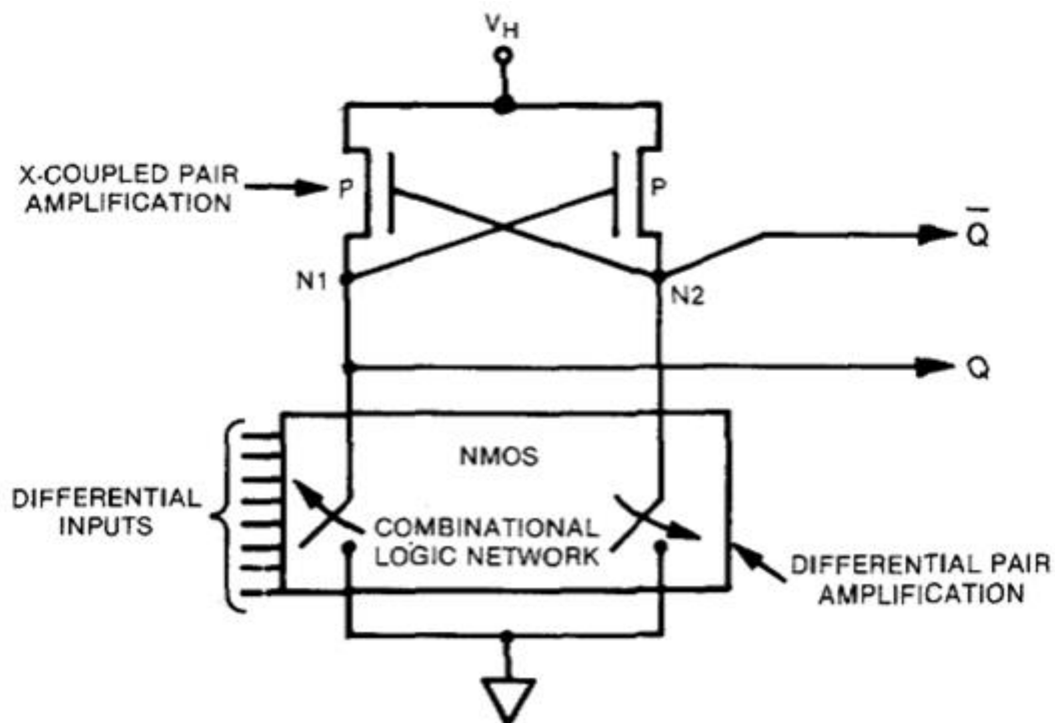


Fig 20: DCVSL logic

According to the figure above, the nodes N1 and N2 are either pulled high or low according to the switching of the inputs. The DCVSL is designed differentially for the given logic in which the true and complementary inputs to the gate produces the complementary outputs . This structure also consumes no static power, like standard CMOS, and also it utilizes the latches in order to generate the output. In this logic style,

large PFETS are eliminated from each logic function, which allows complex circuits to consume low power. A logic function and complement of it is inevitably realized at the same stage . This style can be divided into two basic parts a differential latching circuit and a cascaded complementary logic array . DCVSL is a differential method which requires both true and complementary signals to be routed to gates. Two complementary nMOSFET switching trees are constructed to a pair of cross-coupled pMOSFET transistors. Depending on the differential inputs one of the outputs is pulled down by the corresponding nMOSFET network . The differential output is then latched by the cross-coupled pMOSFET transistors. Since the inputs drive only the nMOSFET transistors of the switching trees, the input capacitance is typically two or three times smaller than that of the conventional static CMOS logic. The advantage of DCVSL is in its logic density that is achieved by elimination of large pMOSFETs from each logic function. Both pull down networks in the Fig. will never conduct at the same time .

2.3.4 TRANSMISSION GATE LOGIC

A **transmission gate** (**TG**) is similar to a relay that can conduct in both directions or block by a control signal with almost any voltage potential. It is a CMOS-based switch, in which PMOS passes a strong 1 but poor 0, and NMOS passes strong 0 but poor 1. Both PMOS and NMOS work simultaneously.

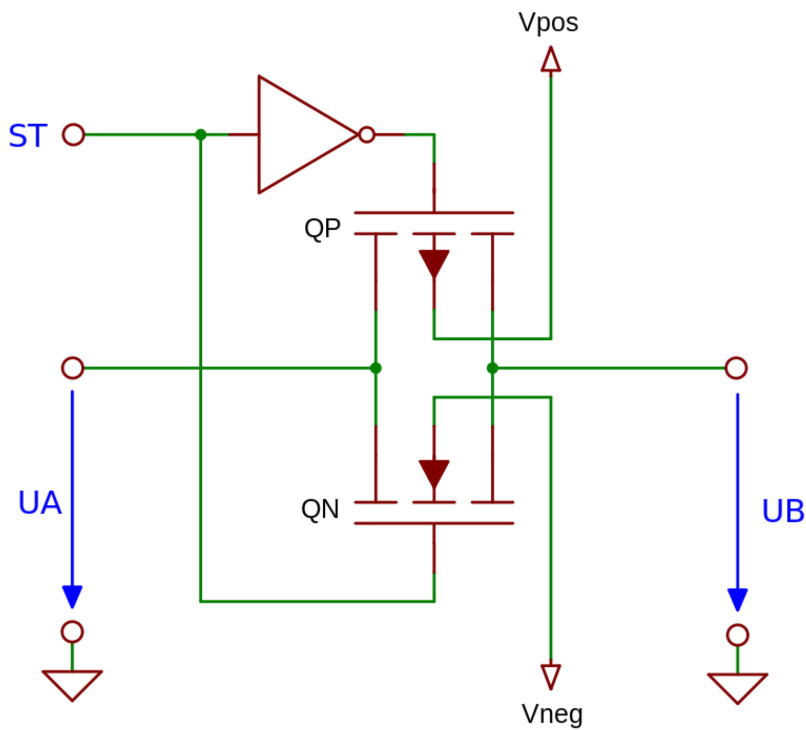


Fig 21: TG Logic

In principle, a transmission gate made up of two field-effect transistors, in which – in contrast to traditional discrete field-effect transistors – the substrate terminal (bulk) is not connected internally to the source terminal. The two transistors, an n-channel MOSFET and a p-channel MOSFET, are connected in parallel with this, however, only the drain and source terminals of the two transistors are connected together. Their gate terminals are connected to each other by a NOT gate (inverter), to form the control terminal.

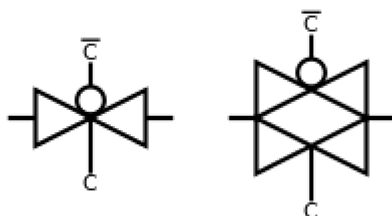


Fig 21: Bowtie symbol of TG

Two variants of the "bow tie" symbol commonly used to represent a transmission gate in circuit diagrams

Unlike with discrete FETs, the substrate terminal is not connected to the source connection. Instead, the substrate terminals are connected to the respective supply potential in order to ensure that the parasitic substrate diode (between gate and substrate) is always reversely biased and so does not affect signal flow. The substrate terminal of the p-channel MOSFET is thus connected to the positive supply potential, and the substrate terminal of the n-channel MOSFET connected to the negative supply potential.

Function of TG:

When the control input is a logic zero (negative power supply potential), the gate of the n-channel MOSFET is also at a negative supply voltage potential. The gate terminal of the p-channel MOSFET is caused by the inverter, to the positive supply voltage potential. Regardless of on which switching terminal of the transmission gate (A or B) a voltage is applied (within the permissible range), the gate-source voltage of the n-channel MOSFETs is always negative, and the p-channel MOSFETs is always positive. Accordingly, neither of the two transistors will conduct and the transmission gate turns off.

When the control input is a logic one, the gate terminal of the n-channel MOSFETs is located at a positive supply voltage potential. By the inverter, the gate terminal of the p-channel MOSFETs is now at a negative supply voltage potential. As the substrate terminal of the transistors is not connected to the source terminal, the drain and source terminals are almost equal and the transistors start at a voltage difference between the gate terminal and one of these conducts.

One of the switching terminals of the transmission gate is raised to a voltage near the negative supply voltage, a positive gate-source voltage (gate-to-drain voltage) will occur at the N-channel MOSFET, and the transistor begins to conduct, and the transmission gate conducts. The voltage at one of the switching terminals of the transmission gate is now raised continuously up to the positive supply voltage potential, so the gate-source voltage

is reduced (gate-drain voltage) on the n-channel MOSFET, and this begins to turn off. At the same time, the p-channel MOSFET has a negative gate-source voltage (gate-to-drain voltage) builds up, whereby this transistor starts to conduct and the transmission gate switches.

Thereby it is achieved that the transmission gate passes over the entire voltage range. The transition resistance of the transmission gate varies depending upon the voltage to be switched, and corresponds to a superposition of the resistance curves of the two transistors.

2.3.5 PASS TRANSISTOR LOGIC:

In electronics, **pass transistor logic** (PTL) describes several **logic** families used in the design of integrated circuits. It reduces the count of **transistors** used to make different **logic** gates, by eliminating redundant **transistor.**

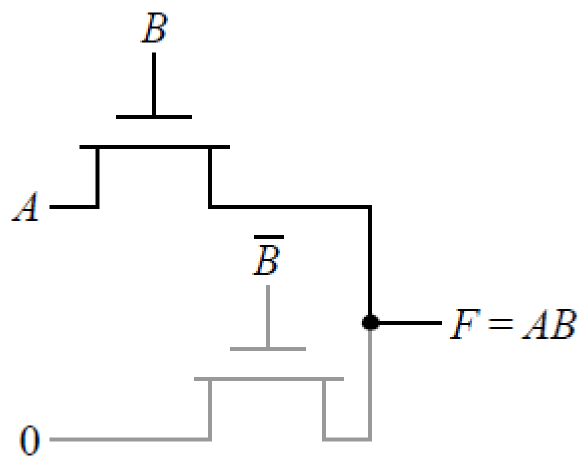


Fig 22: Pass Transistor Logic

When B is “1”, top device turns on and copies the input A to output F. When B is low, bottom device turns on and passes a “0”. The presence of the switch driven by B is essential to

ensure that the gate is static – a low-impedance path must exist to supply rails.

Adv.: Fewer devices to implement some functions.Example: AND2 requires 4 devices

(including inverter to invert B) vs. 6 for complementary CMOS (lower total capacitance). NMOS is effective at passing a 0, but poor at pulling a node to Vdd. When the pass transistor a

node high, the output only charges up to Vdd-Vtn. This becomes worse due to the body effect.

The node will be charged up to Vdd – Vtn (Vs).

Pass transistors require **lower switching energy** to charge up a node, due to the reduces voltage swing. The output node charges from 0 -> Vdd-Vtn, and the energy drawn from the

power supply for charging the output of a pass transistor is given by CL.Vdd(Vdd-Vtn)While lower switching power is consumed, it may consume static power when output is high –the reduced voltage level may be insufficient to turn off the PMOS transistor of the subsequent

CMOS inverter.

2.3.6 COMPLEMENTARY PASS TRANSISTOR LOGIC

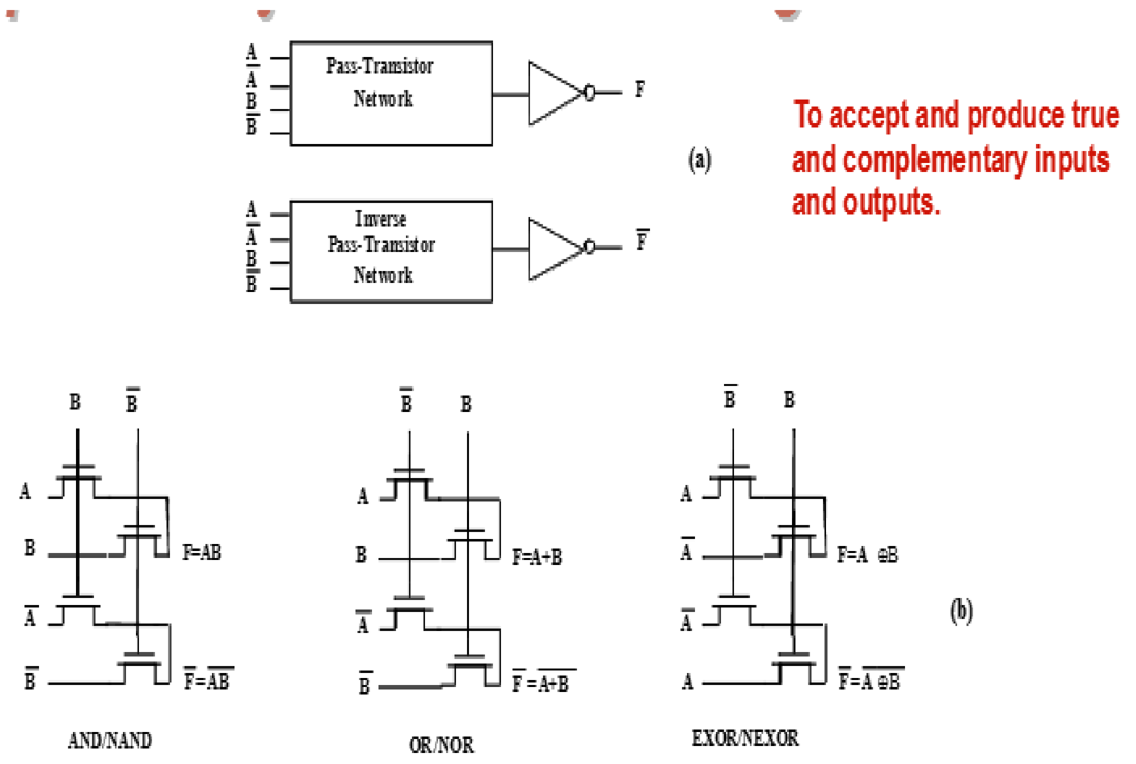


Fig 23: CPL Logic

Since circuit is differential, complimentary inputs and outputs are available. Although generating differential signals require extra circuitry, complex gates such as XORs, MUXs and adders can be realized efficiently.

•CPL is a static gate, because outputs are connected to Vdd or GND through a low-resistance path (high noise resilience).

•Design is modular all gates use same topology; only inputs are permuted. This fac ilitates the design of a library of gates.

2.3.7 DYNAMIC LOGIC

In static circuits at every point in time (except when switching) the output is connected to either GND or VDD via a low resistance path. fan-in of N requires 2N devices Dynamic circuits rely on the temporary storage of signal values on the capacitance of high impedance nodes.

** requires only N + 2 transistors takes a sequence of precharge and conditional evaluation**

phases to realize logic functions.

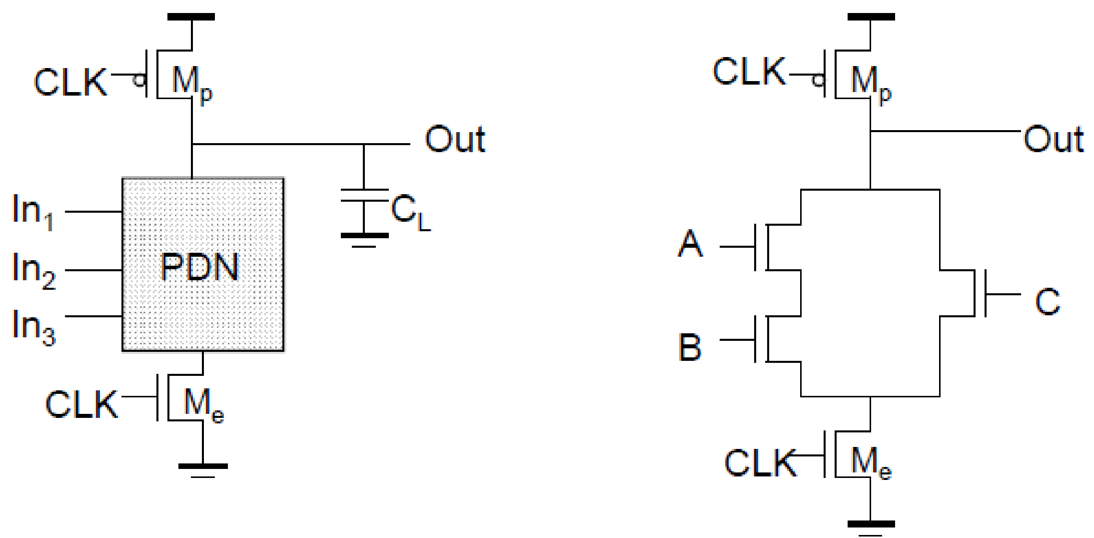


Fig 24: Dynamic Logic

Once the output of a dynamic gate is discharged, it cannot be charged again until the next precharge operation. Inputs to the gate can make at most one transition during evaluation. Output can be in the high impedance state during and after evaluation (PDN off), state is stored on CL.

Logic function is implemented by the PDN only number of transistors is N + 2 (versus 2N for static complementary CMOS) should be smaller in area than static complementary CMOS

Full swing outputs (VOL = GND and VOH = VDD)

Nonratioed - sizing of the devices is not important for proper functioning (only for performance)

- Faster switching speeds
- reduced load capacitance due to lower number of transistors per gate (Cint) so a reduced logical effort

- reduced load capacitance due to smaller fan-out (Cext)
- no Isc, so all the current provided by PDN goes into discharging CL
- Ignoring the influence of precharge time on the switching speed of

the gate, tpLH = 0 but the presence of the evaluation transistor slows down the tpHL

Power dissipation should be better

- consumes only dynamic power – no short circuit power
  consumption since the pull-up path is not on when evaluating
- lower CL- both Cint (since there are fewer transistors connected to
  the drain output) and Cext (since there the output load is one per
  connected gate, not two)
- by construction can have at most one transition per cycle – no
- glitching
- But power dissipation can be significantly higher due to
- higher transition probabilities
  extra load on CLK
- PDN starts to work as soon as the input signals exceed
- VTn, so set VM, VIH and VIL all equal to VTn
- low moise margin (NML)
- Needs a precharge clock

2.3.8 CASCADING PROBLEM IN DYNAMIC CMOS LOGIC

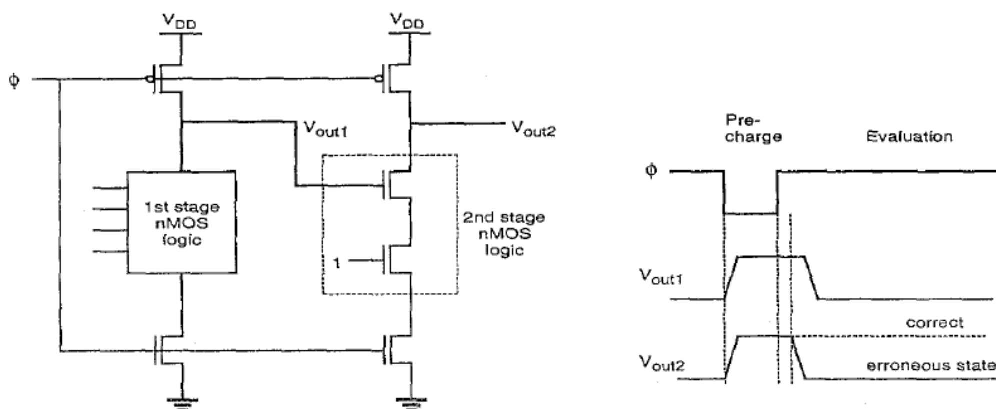


Fig 25: Cascading in Dynamic CMOS

If several stages of the previous CMOS dynamic logic circuit are cascaded together using the same clock ϕ, a problem in evaluation involving a built -in “race condition” will exist

•Consider the two stage dynamic logic circuit below:

During pre-charge, both Vout1 and Vout2 are pre-charged to Vdd

–When ϕ goes high to begin evaluate, all inputs at stage 1 require some finite time to resolve,but during this time charge may erroneously be discharged from Vout2

•e.g. assume that eventually the 1st stage NMOS logic tree conducts and fully discharges Vout1, but since all the inputs to the N-tree all not immediately resolved, it takes some time for the N-tree to finally discharge Vout1 to GND.

•If, during this time delay, the 2nd stage has the input condition shown with bottom NMOS transistor gate at a logic 1, then Vout2 will start to fall and discharge its load capacitance until Vout1 finally evaluates and turns off the top series NMOS transistor in stage 2

The result is an error in the output of the 2nd stage Vout2.

With simple cascading of dynamic CMOS logic stages, a problem arises in

the evaluate cycle:

– The pre-charged high voltage on Node N2 in stage 2 may be inadvertently (partially) discharged by logic inputs to stage 2 which have not yet reached final correct (low) values from the stage 1 evaluation operation.
– Can not simply cascade dynamic CMOS logic gates without preventing unwanted bleeding of charge from pre-charged nodes
• Possible Solutions:

– two phase clocks

– use of inverters to create Domino Logic

– NP Domino Logic

– Zipper/NORA logic

2.3.9 DOMINO Logic

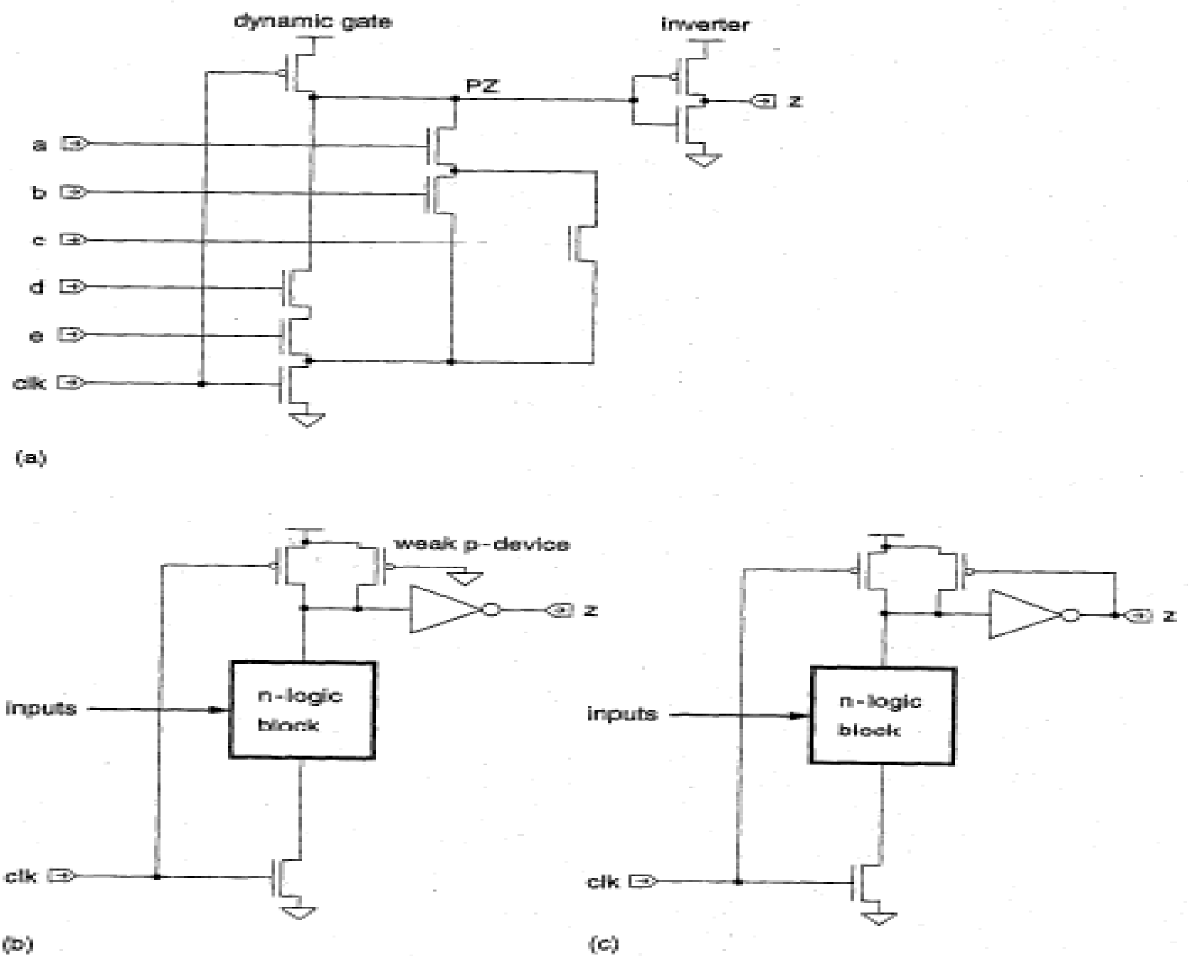


Fig 26:Domino logic

The problem with faulty discharge of precharged nodes in CMOS dynamic logic circuits can be solved by placing an inverter in series with the output of each gate

– All inputs to N logic blocks (which are derived from inverted outputs of previous stages) therefore will be at zero volts during precharge and will remain at zero until the evaluation stage has logic inputs to discharge the precharged node PZ.

– This circuit approach avoids the race problem of "vanilla" cascaded dynamic CMOS

– However, all circuits only provide noninverted outputs • In (b) a weak P device compensates for charge loss due to charge sharing and leakage at low frequency clock operation

• In (c) the weak P device can be used to latch the output high

NORA LOGIC

An elegant solution to the dynamic CMOS logic “erroneous evaluation” problem is to

use NP Domino Logic (also called NORA logic) as shown below.

– Alternate stages of N logic with stages of P logic

• N logic stages use true clock, normal precharge and evaluation phases, with N logic tree in the pull down leg. P logic stages use a complement clock, with P logic stage tied above the output node.

• During precharge clk is low (-clk is high) and the P-logic output precharges to ground while N-logic outputs precharge to Vdd.

• During evaluate clk is high (-clk is low) and both type stages go through evaluation; N-logic tree logically evaluates to ground while P-logic tree logically evaluates to Vdd.

• Inverter outputs can be used to feed other N-blocks from N-blocks, or to feed other Pblocks from P-blocks

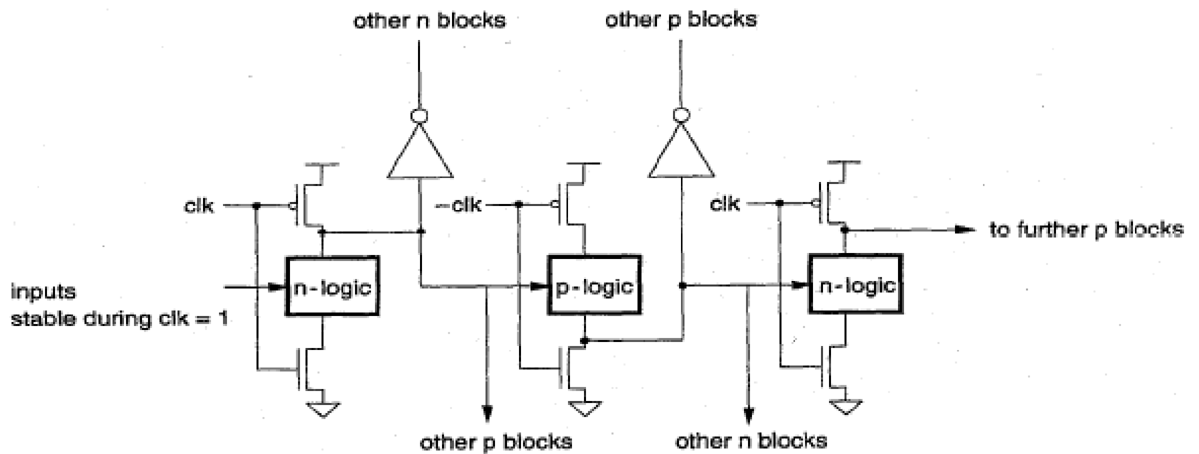


Fig 27:

2.3.10 NORA CMOS Logic Circuit Example

An example of NP or NORA (No Race) logic is shown below:

• During ϕ low (ϕ' high), each stage pre-charges

– N logic stages pre-charge to Vdd; P logic stages pre-charge to GND

• When ϕ**,** goes high ϕ' low), each stage enters the evaluation phase

– N logic evaluates to GND; P logic stages evaluate to Vdd

– All NMOS and PMOS stages evaluate one after another in succession, as in Domino logic

• Logic below:

– Stage 1 is X = (A · B)'

– Stage 2 is G = X' + Y'

– Stage 3 is Z = (F · G + H)'

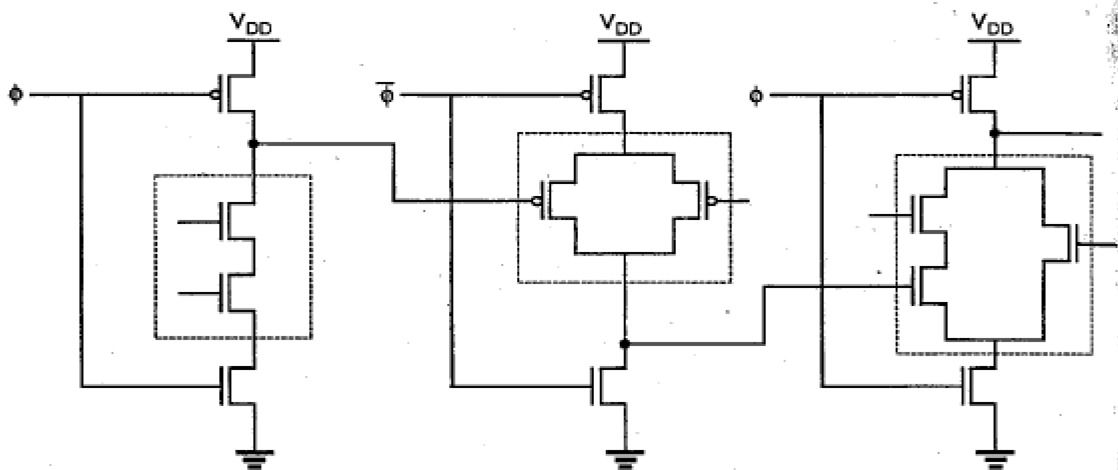


Fig 28: NORA

2.4.Sequential Circuit and Semiconductor Memory Design.

2.4.1Sequential Circuit and Semiconductor Memory Design:

Logic circuits are divided into two categories − (a) Combinational Circuits, and (b) Sequential Circuits.

In Combinational circuits, the output depends only on the condition of the latest inputs.

In Sequential circuits, the output depends not only on the latest inputs, but also on the condition of earlier inputs. Sequential circuits contain memory elements.

Sequential circuits are of three types −

Bistable − Bistable circuits have two stable operating points and will be in either of the states. Example − Memory cells, latches, flip-flops and registers.

Monostable − Monostable circuits have only one stable operating point and even if they are temporarily perturbed to the opposite state, they will return in time to their stable operating point. Example: Timers, pulse generators.

Astable − circuits have no stable operating point and oscillate between several states. Example − Ring oscillator.

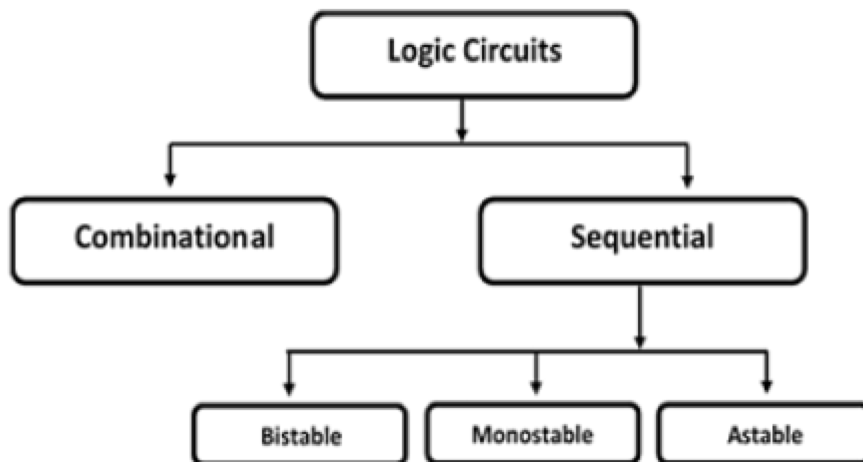


Fig 29 : Classification of Sequential Logic

2.4.2SR Latch based on NOR Gate

If the set input (S) is equal to logic **"1"** and the reset input is equal to logic **"0."** then the output Q will be forced to logic **"1"**. While $\bar{Q}$ is forced to logic **"0"**. This means the SR latch will be set, irrespective of its previous state.
Similarly, if S is equal to "0" and R is equal to **"1"** then the output Q will be forced to **"0"** while Q and $\bar{Q}$ is forced to **"1"**. This means the latch is reset, regardless of its previously held state. Finally, if both of the inputs S and R are equal to logic **"1"** then both output will be forced to logic **"0"** which conflicts with the complementarity of Q and $\bar{Q}$.
Therefore, this input combination is not allowed during normal operation. Truth table of NOR based SR Latch is given in table.

S	R	Q	$\overline{Q}$	Operation
0	0	Q	$\overline{Q}$	Hold
1	0	1	0	Set
0	1	0	1	Reset
1	1	0	0	Not allowed

CMOS SR latch based on NOR gate is shown in the figure given below.

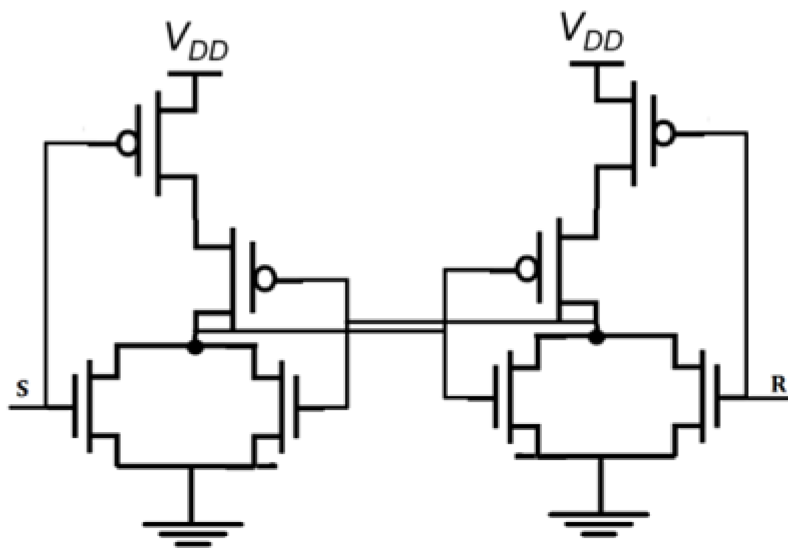


Fig 30: CMOS SR Latch

If the S is equal to V_{OH} and the R is equal to V_{OL}, both of the parallel-connected transistors M1 and M2 will be ON. The voltage on node $\overline{Q}$ will assume a logic-low level of $V_{OL} = 0$.

At the same time, both M3 and M4 are turned off, which results in a logic-high voltage V_{OH} at node Q. If the R is equal to V_{OH} and the S is equal to V_{OL}, M1 and M2 turned off and M3 and M4 turned on.

SR Latch based on NAND Gate

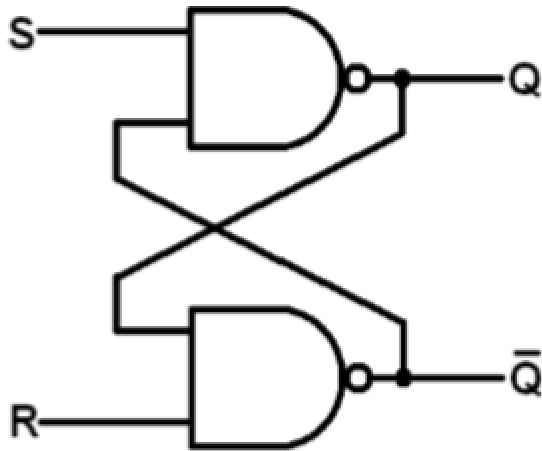


Fig 31: NAND Latch

Block diagram and gate level schematic of NAND based SR latch is shown in the figure. The small circles at the S and R input terminals represents that the circuit responds to active low input signals. The truth table of NAND based SR latch is given in table.

S	R	Q	Q'	
0	0	NC	NC	No change. Latch remained in present state.
1	0	1	0	Latch SET.
0	1	0	1	Latch RESET.
1	1	0	0	Invalid condition.

If S goes to 0 (while R = 1), Q goes high, pulling $\bar{Q}$ low and the latch enters Set state

S = **0** then Q = **1** (if R = **1**)

If R goes to 0 (while S = 1), Q goes high, pulling $\bar{Q}$ low and the latch is Reset

R = **0** then Q = **1** (if S = **1**)

Hold state requires both S and R to be high. If S = R = 0 then output is not allowed, as it would result in an indeterminate state. CMOS SR Latch based on NAND Gate is shown in figure.

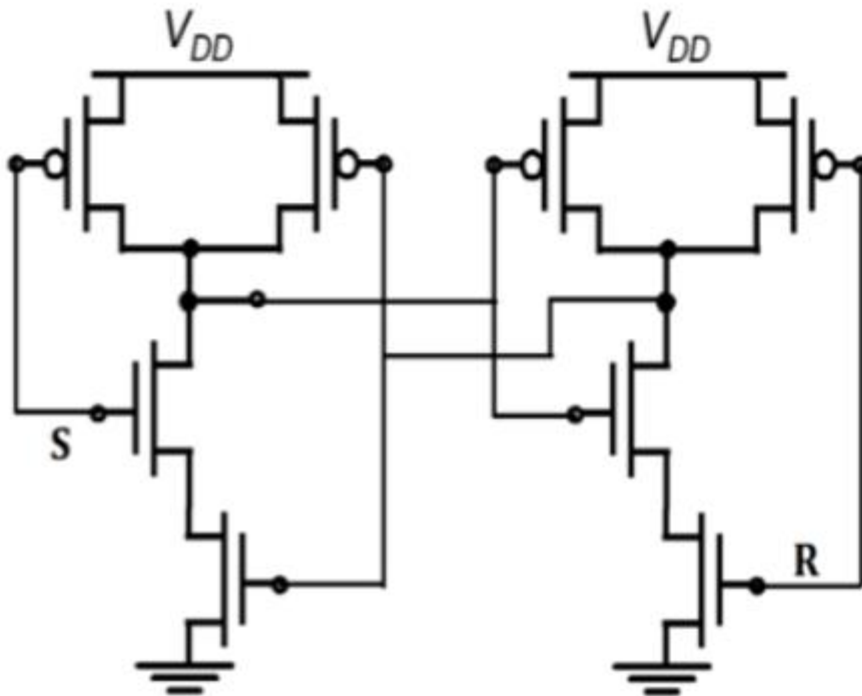


Fig 32: CMOS SR Latch based on NAND Gate

Depletion-load nMOS SR Latch based on NAND Gate is shown in figure. The operation is similar to that of CMOS NAND SR latch. The CMOS circuit implementation has low static power dissipation and high noise margin.

Clocked SR Latch based on NAND Gate

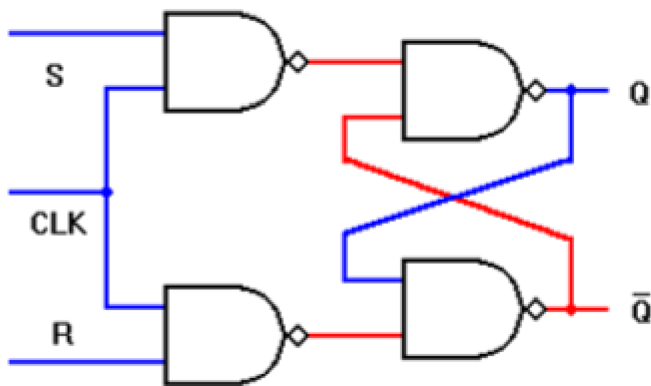


Fig 33: Clocked SR Latch based on NAND gate

Circuit is implemented with four NAND gates. If this circuit is implemented with CMOS then it requires 16 transistors.

- The latch is responsive to S or R only if CLK is high.
- If both input signals and the CLK signals are active high: i.e., the latch output Q will be set when CLK = "1" S = "1" and R = "0"
- Similarly, the latch will be reset when CLK = "1," S = "0," and

When CLK is low, the latch retains its present state.

2.4.3Clocked JK Latch

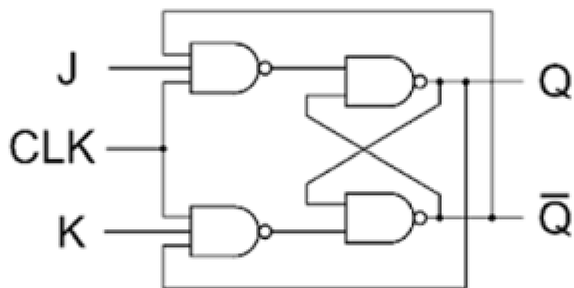


Fig 34: Clocked JK

The figure above shows a clocked JK latch, based on NAND gates. The disadvantage of an SR latch is that when both S and R are high, its output state becomes indeterminant. The JK latch eliminates this problem by using feedback from output to input, such that all input states of the truth table are allowable. If J = K = 0, the latch will hold its present state.If J = 1 and K = 0, the latch will set on the next positive-going clock edge, i.e. Q = 1, $\bar{Q}$ = 0.If J = 0 and K = 1, the latch will reset on the next positive-going clock edge,

If J = K = 1, the latch will toggle on the next positive-going clock edge.

The operation of the clocked JK latch is summarized in the truth table given in table.

J	K	Q	$\overline{Q}$	S	R	Q	$\overline{Q}$	Operation
0	0	0	1	1	1	0	1	Hold
		1	0	1	1	1	0	
0	1	0	1	1	1	0	1	Reset
		1	0	1	0	0	1	
1	0	0	1	0	1	1	0	Set
		1	0	1	1	1	0	
1	1	0	1	0	1	1	0	toggle
		1	0	1	0	0	1	

2.4.4CMOS D Latch Implementation

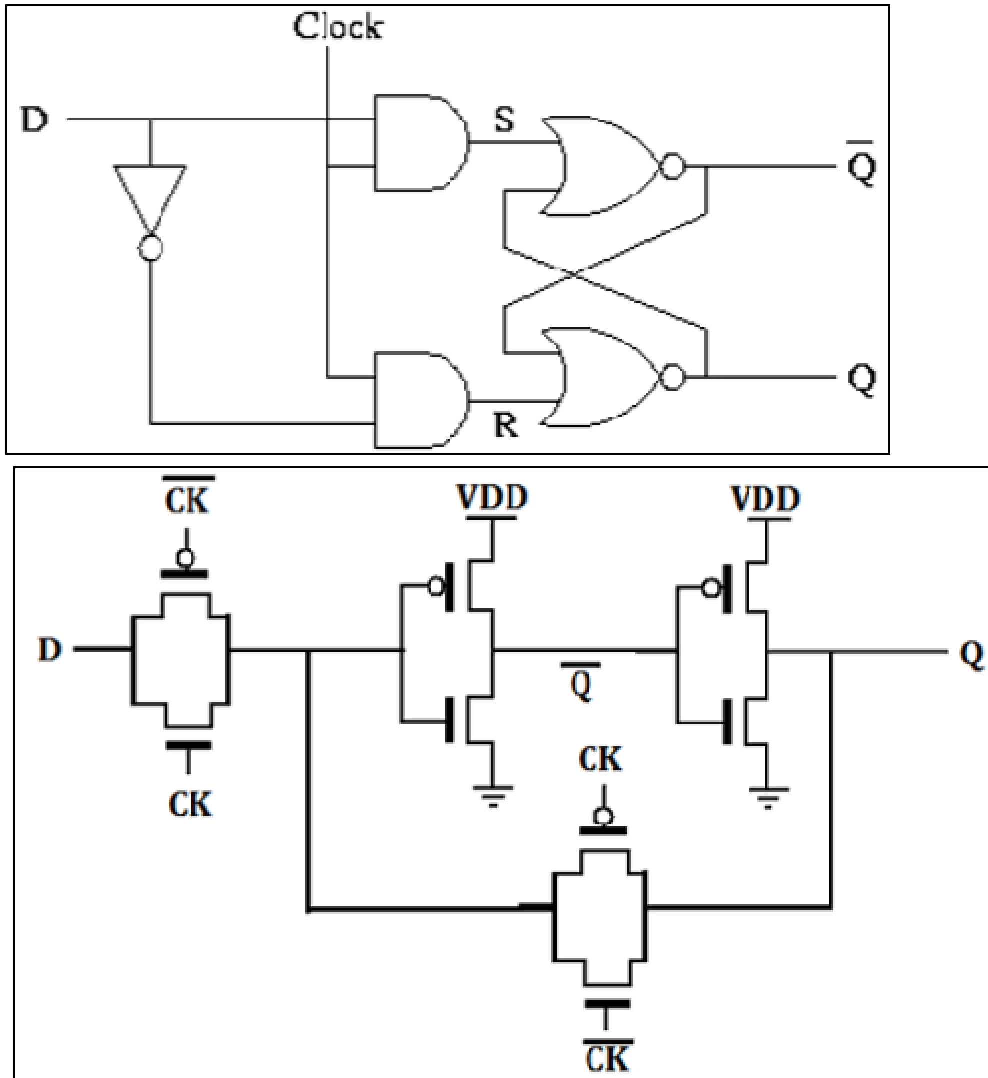


Fig 35: D latch using TG

The D latch is normally, implemented with transmission gate (TG) switches as shown in the figure. The input TG is activated with CLK while the latch feedback loop TG is activated with CLK. Input D is accepted when CLK is high. When CLK goes low, the input is open circuited and the latch is set with the prior data D.

2.4.5 SRAM Basics: The memory circuit is said to be static if the stored data can be retained indefinitely, as long as the power supply is on, without any need for periodic refresh operation. The data storage cell, i.e., the one-bit memory cell in the static RAM arrays, invariably consists of a simple latch circuit with two stable operating points. Depending on the preserved state of the two inverter latch circuit, the data being held in the memory cell will be interpreted either as logic '0' or as logic '1'. To access the data contained in the memory cell via a bit line, we need atleast one switch, which is controlled by the corresponding word line as shown in Figure below.

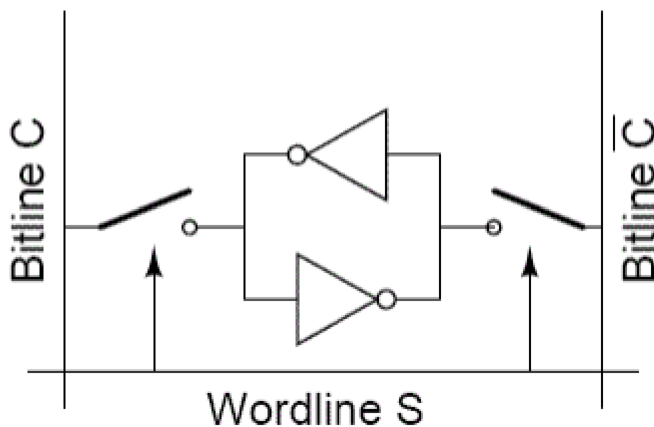


Fig 36: SRAM Basics

CMOS SRAM Cell :A low power SRAM cell may be designed by using cross-coupled CMOS inverters. The most important advantage of this circuit topology is that the static power dissipation is very small; essentially, it is limited by small leakage current. Other advantages of this design are high noise immunity due to larger noise margins, and the ability to operate at lower power supply voltage. The major disadvantage of this topology is larger cell size. The circuit structure of the full CMOS static RAM cell is shown in Figure 28.12. The memory cell consists of simple CMOS inverters connected back to

back, and two access transistors. The access transistors are turned on whenever a word line is activated for read or write operation, connecting the cell to the complementary bit line columns.

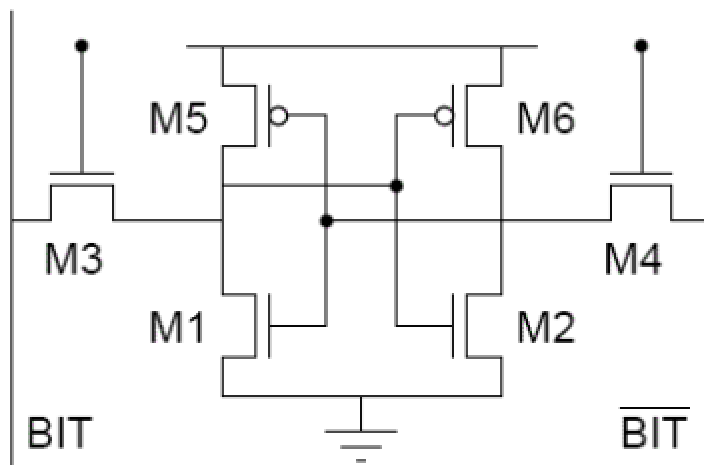


Fig 37: 6T SRAM

CMOS SRAM Cell Design To determine W/L ratios of the transistors, a number of design criteria must be taken into consideration. The two basic requirements, which dictate W/L ratios, are that the data read operation should not destroy the stored information in the cell. The cell should allow stored information modification during write operation. In order to consider operations of SRAM, we have to take into account, the relatively large parasitic column capacitance Cbit and C$\overline{bit}$ and column pull-up transistors as shown in Figure 38.

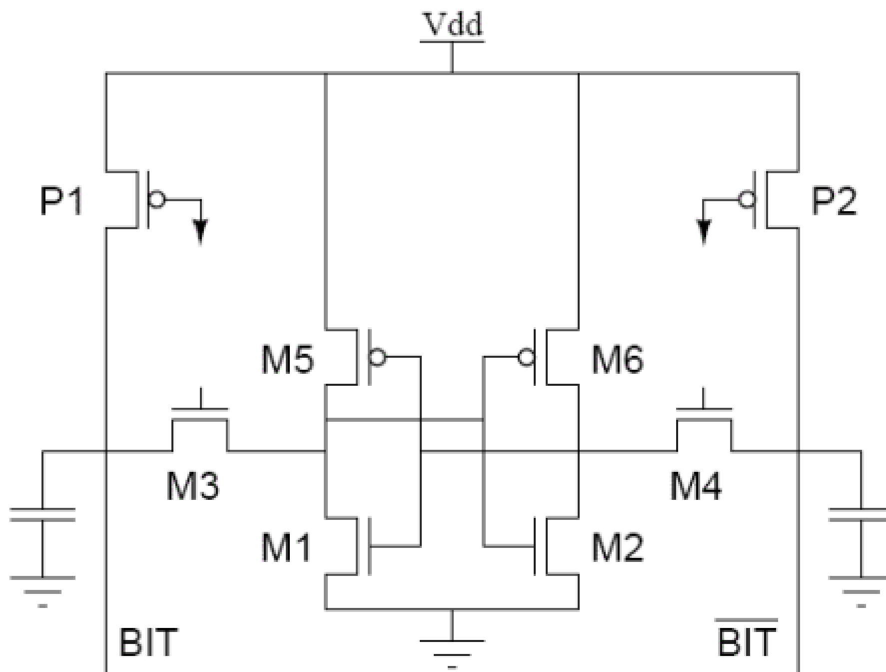


Fig 38: 6T SRAM with capacitance

When none of the word lines is selected, the pass transistors M3 and M4 are turned off and the data is retained in all memory cells. The column capacitances are charged by the pull-up transistors P1 and P2. The voltages across the column capacitors reach VDD - VT.

READ Operation Consider a data read operation, shown in Figure below, assuming that logic '0' is stored in the cell. The transistors M2 and M5 are turned off, while the transistors M1 and M6 operate in linear mode. Thus internal node voltages are V1 = 0 and V2 = VDD before the cell access transistors are turned on. The active transistors at the beginning of data read operation are shown in Figure

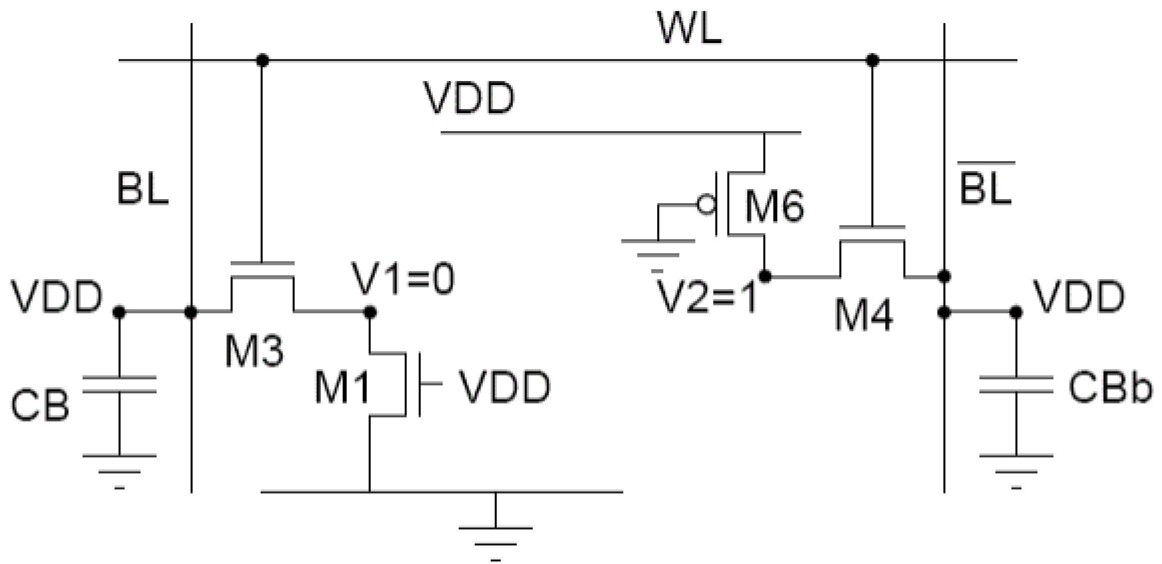


Fig 39: READ in 6T SRAM

After the pass transistors M3 and M4 are turned on by the row selection circuitry, the voltage CBb of will not change any significant variation since no current flows through M4. On the other hand M1 and M3 will conduct a nonzero current and the voltage level of CB will begin to drop slightly. The node voltage V1 will increase from its initial value of '0'V. The node voltage V1 may exceed the threshold voltage of M2 during this process, forcing an unintended change of the stored state. Therefore voltage must not exceed the threshold voltage of M2, so the transistor M2 remains turned off during read phase. The transistor M3 is in saturation whereas M1 is linear.

WRITE Operation Consider the write '0' operation assuming that logic '1' is stored in the SRAM cell initially. Figure 28.51 shows the voltage levels in the CMOS SRAM cell at the beginning of the data write operation. The transistors M1 and M6 are turned off, while M2 and M5 are operating in the linear mode. Thus the internal node voltage V1 = VDD and V2 = 0 before the access transistors are turned on. The column voltage Vb is forced to '0' by the write circuitry. Once M3 and M4 are turned on, we expect the nodal

voltage V2 to remain below the threshold voltage of M1, since M2 and M4 are designed according to

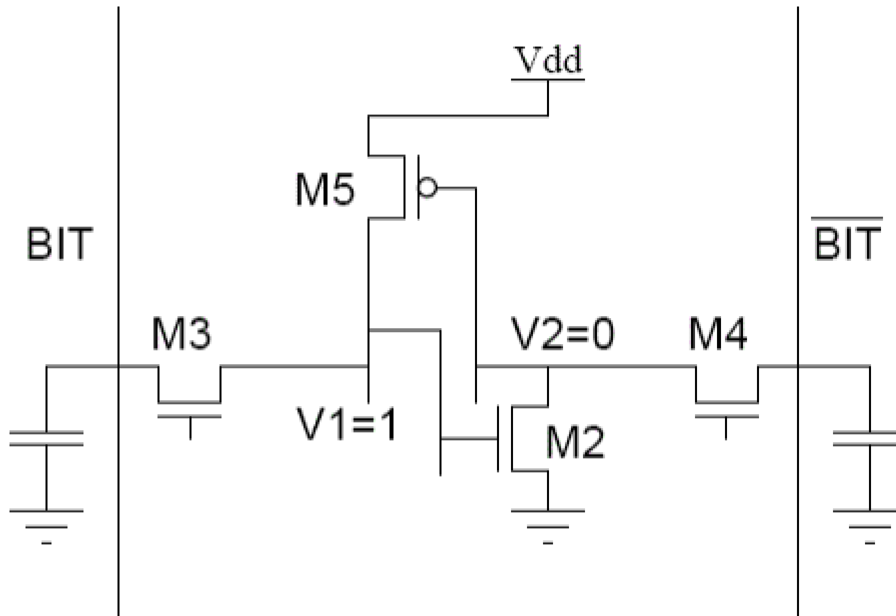


Fig 40:WRITE In 6T SRAM

The voltage at node 2 would not be sufficient to turn on M1. To change the stored information, i.e., to force V1 = 0 and V2 = VDD, the node voltage V1 must be reduced below the threshold voltage of M2, so that M2 turns off. When the transistor M3 operates in linear region while M5 operates in saturation region.

2.4.6 3T DRAM Cell:

The simplest DRAM cell is the 3T scheme. A 3T DRAM cell has a higher density than a SRAM cell; moreover in a 3T DRAM, there is no constraint on device ratios and the read operation is nondestructive. In this cell, the storage capacitance is the gate capacitance of the readout device, so making this scheme attractive for embedded memory applications; however, a 3T DRAM shows still limited performance and low retention time to severely limit its use in advanced integrated circuits.3T DRAM utilizes gate of the transistor and

a capacitance to store the data value. When data is to be written, write signal is enabled and the data from the bit line is fed into the cell. When data is to be read from the cell, read line is enabled and data is read through the bit line.

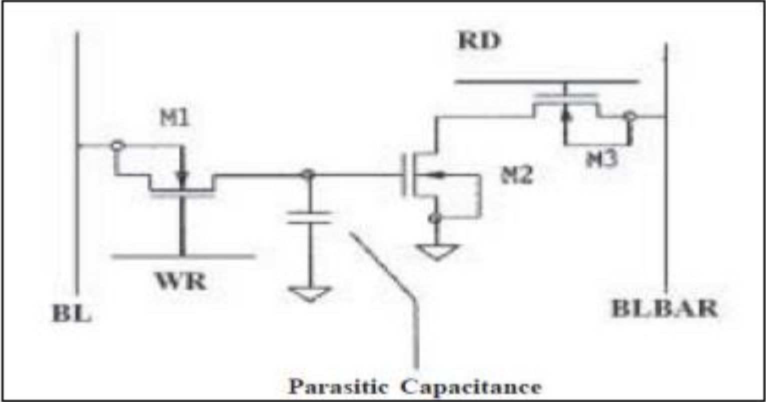


Fig 41: 3T DRAM

Module 3 : Analog VLSI Circuit Design

3.1 Small signal model of n- channel MOSFET

The model of small signal model is shown in fig. 3.1

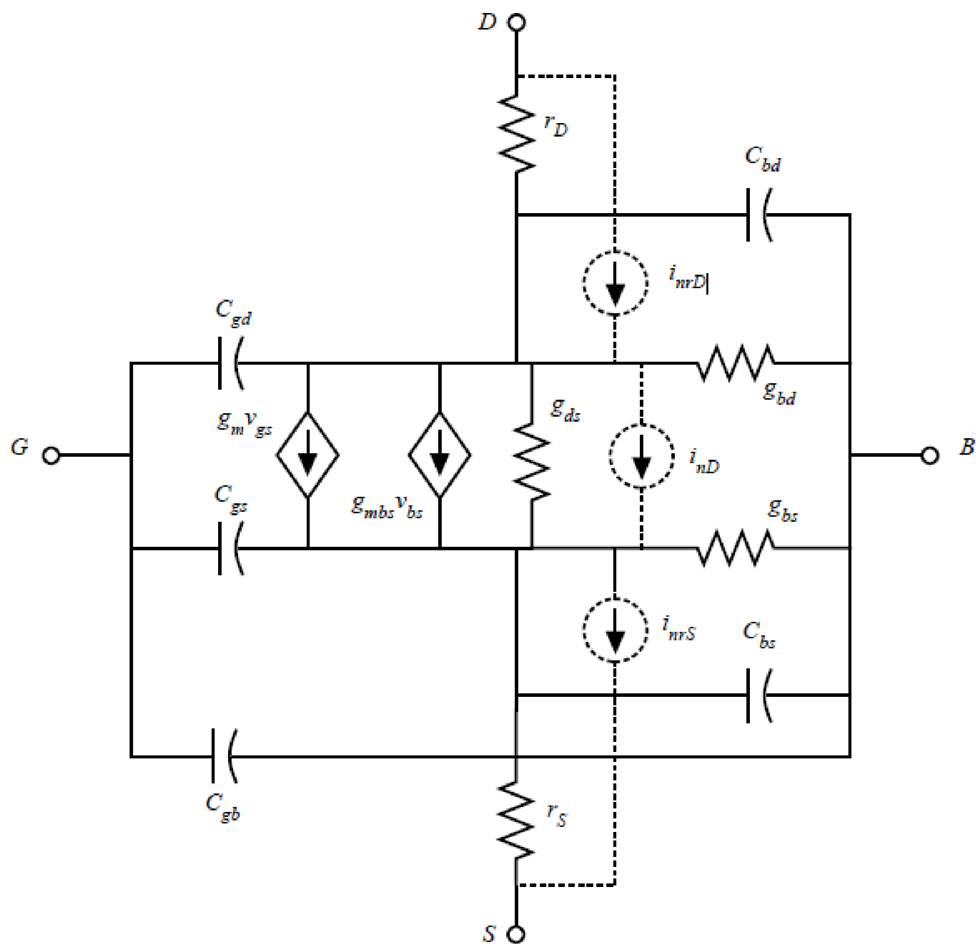


Fig.3.1 : Small signal model of NMOS

It includes internal capacitances, resistances, controlled current sources.

3.2 Analog sub-circuits

Analog sub-circuits are discussed in this section

3.2.1 MOS Switch

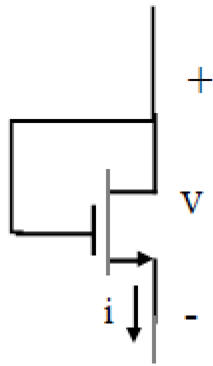
MOS transistor can be used as a controlled switch . Turn on and off is controlled by the voltage applied at gate terminal . An NMOS is turn on when $V_{GS} > V_{TH}$ and $V_{DS} < (V_{GS} - V_{TH})$. The on resistance of the MOS transistor is shown as follows .

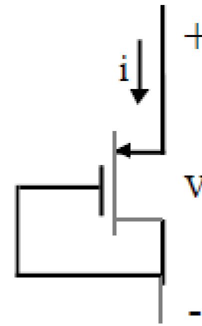
$$R_{ON} = \frac{\partial v_{DS}}{\partial i_D} = \frac{1}{\frac{K'W}{L}(v_{GS} - V_T - v_{DS})} \quad .....(3.1)$$

NMOS is turn off when $V_{GS} < V_{TH}$.

3.2.2 Active resistor and MOS diode

The equivalent circuit of active resistor using MOS transistor is shown in fig. 3.2 a and 3.2b .





3.2a : Active Resistor / MOS diode (NMOS)   3.2b : Active resistor /MOS Diode

3.2.3 Current source and Sink ,Current Mirror

Characterisation of Current source and Sink :

(a) Minimum voltage (V_{MIN}) across current sink or source for which the current is no longer constant.

(b) Output resistance which is a measure of the "flatness" of the current sink or source

The diagram of current sink and source is shown in figure

Current Sink :

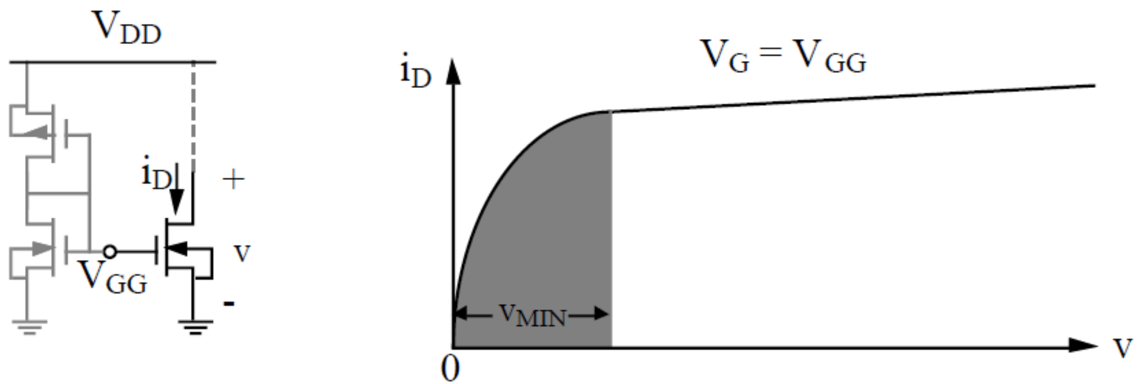
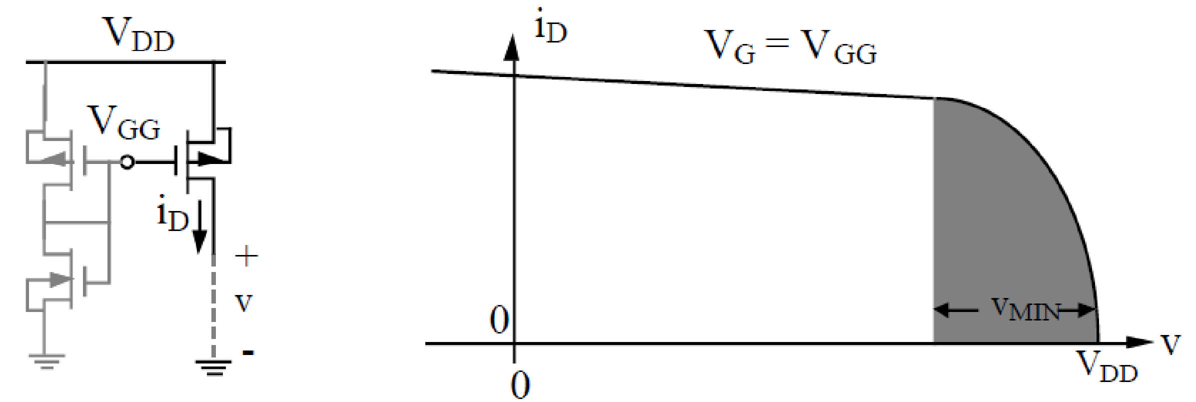


Fig .3.3 : Current Sink

Current Source:



$$r_{OUT} = \frac{1}{\lambda I_D}$$

$v_{MIN} = v_{DS(SAT.)} = v_{ON}$ where $v_{ON} = v_{GS} - V_T$

Fig. 3.4 : Current Source

How to improve the output resistance of current source?

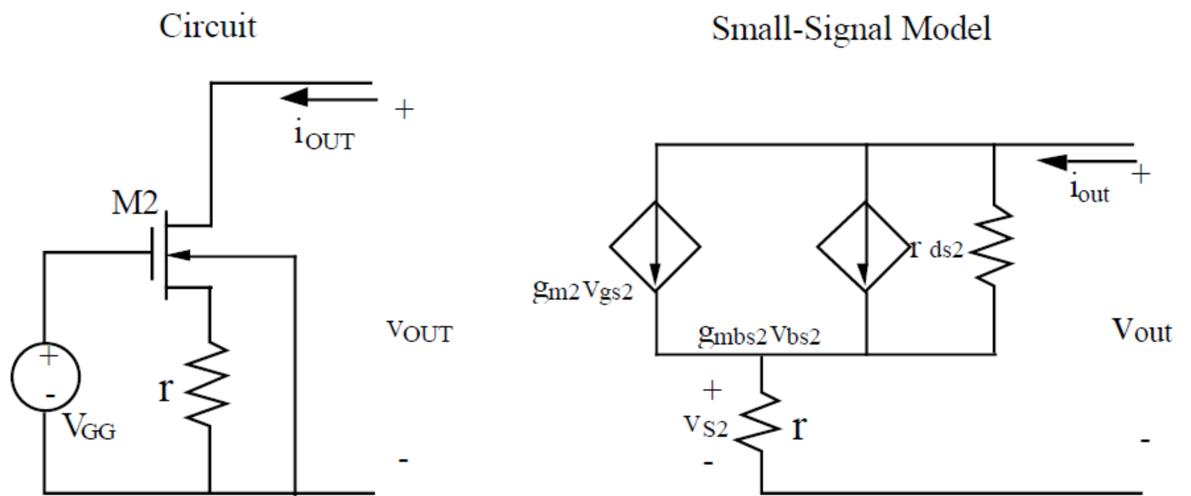


Fig. 3.5 : Increasing of output resistance

Output resistance has been increased as $r_{out} = r_{ds2} + r\,[1 + g_{m2}r_{ds2} + g_{mbs2}r_{ds2}]$

3.2.4 Current and voltage references

A reference circuit is an independent voltage or current source which has a high degree of precision and stability.

Following are desirable from reference source

a) Output voltage/current should be independent of power supply

b) Output voltage/current should be independent of temperature

c) Output voltage/current should be independent of processing variations

V-I Characteristics of an Ideal Reference

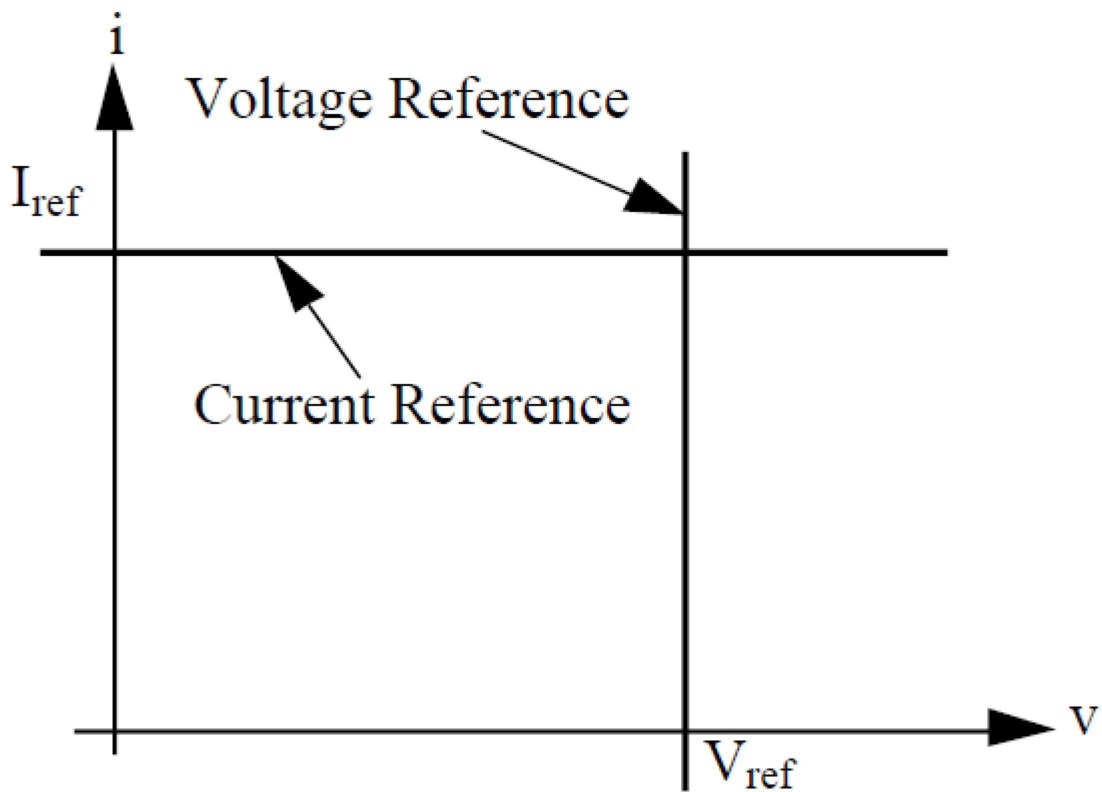


Fig. 3.6 : V-I curve of reference source

Voltage divider

Passive voltage divider is shown in the following figure

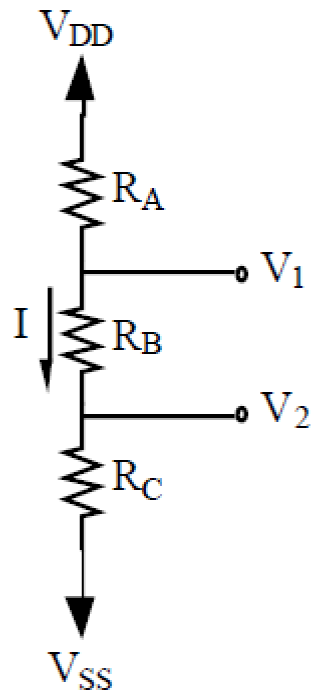


Fig. 3.7 : Passive voltage divider

Active voltage divider circuit is shown in Fig. 3.8

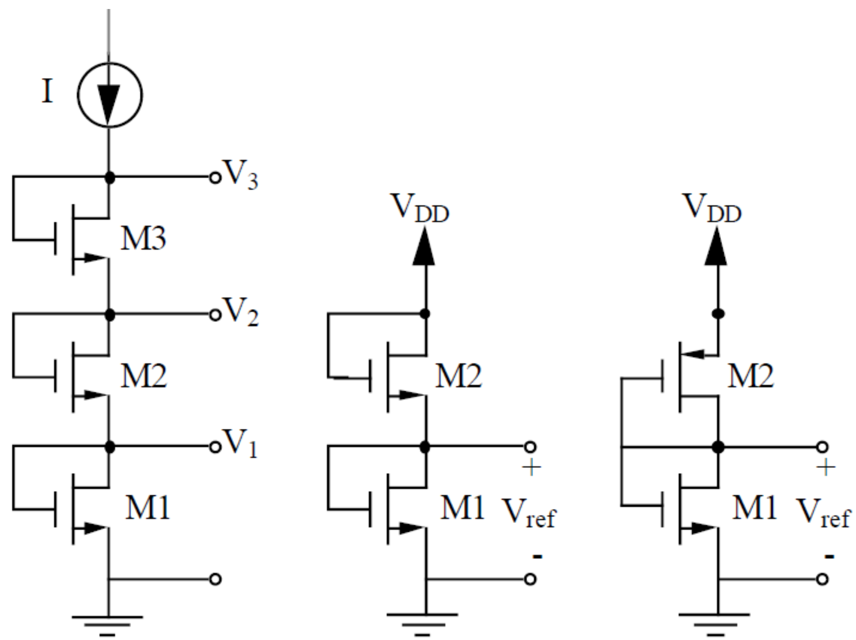


Fig. 3.8 : Active voltage divider circuit

P-N junction Voltage reference

P-N junction voltage reference is shown in Fig. 3.9

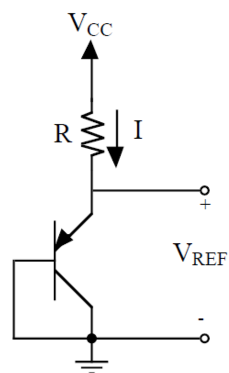


Fig. 3.9 : P-N junction voltage reference

V_{REF} is found as

$$V_{REF} \approx V_t \ln\left(\frac{V_{CC}}{RI_s}\right)$$

Threshold voltage reference (Gate-Source Referenced Circuits / MOS equivalent of the pn junction referenced circuit)

MOS equivalent p-n junction reference is shown in Fig. 3.10

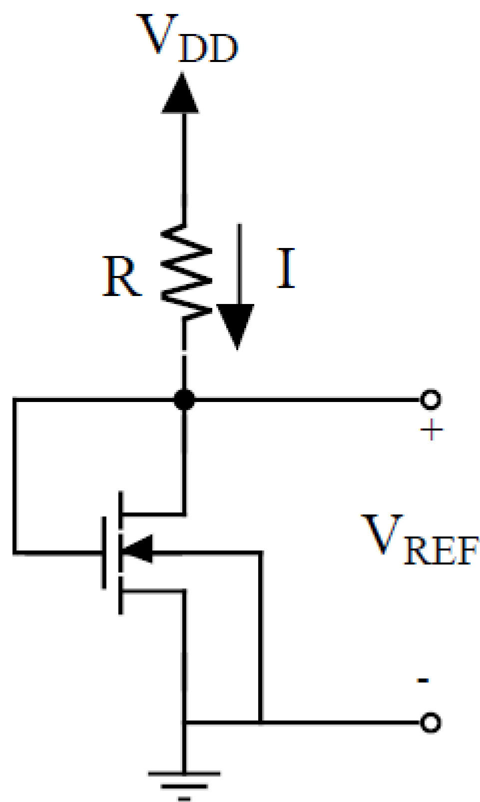


Fig. 3.10: MOS equivalent p-n junction reference

$$V_{REF} = V_T - \frac{1}{\beta R} + \sqrt{\frac{2(V_{DD}-V_T)}{\beta R} + \frac{1}{\beta^2 R^2}}$$

Band gap reference (Basic Principle)

The basic principle of band gap reference is shown in Fig.3.11

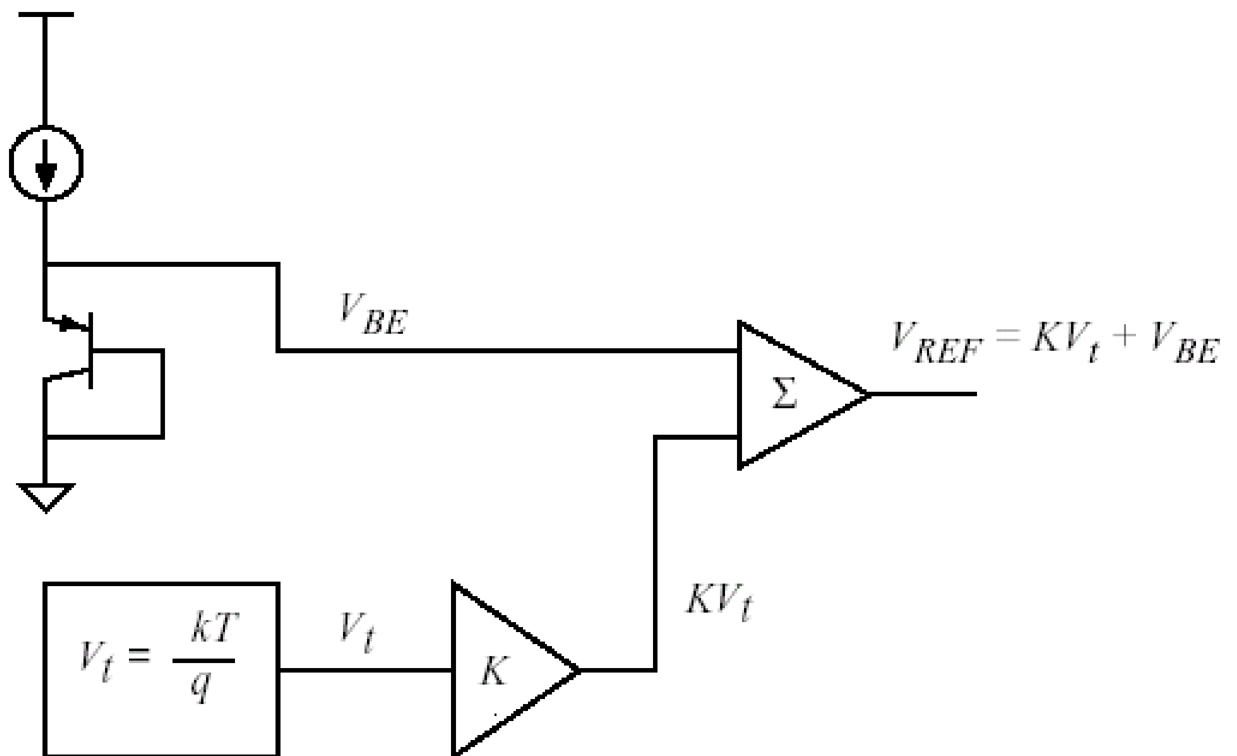


Fig. 3.11: Band gap reference voltage

3.4 Switch-Capacitor Circuit – resistance emulation of series , parallel and series-parallel circuit

- A switched capacitor circuit is an electronic circuit useful for signal processing system
- It works by transferring charge into and out of a capacitor when switches are opened and closed.

- Usually, non-overlapping clock signals are used to control the switches,
- The primary advantage of switched capacitor filters is that they can be easily implemented on an integrated circuit , miniature and low power consumption

In the integrated circuit the resistor can be emulated with the help of MOS switch and capacitors.

The following table summarises the resistor emulation using switch capacitor circuit .

Switched Capacitor Resistor Emulation Circuit	Schematic	Equivalent Resistance
Parallel		$\frac{T}{C}$
Series		$\frac{T}{C}$
Series-Parallel		$\frac{T}{C_1 + C_2}$
Bilinear		$\frac{T}{4C}$

3.5 Switch capacitor integrator and filter (1st order only)

Switch capacitor Integrator

The circuit diagram of integrator circuit is shown in Fig.3.12

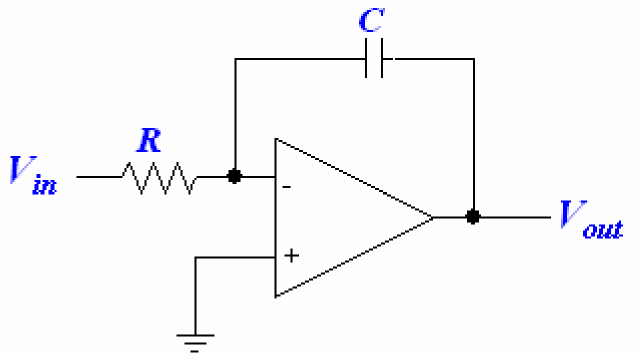


Fig. 3.12 : Circuit diagram of integrator circuit using op-amp

In the circuit of Fig. 3.12 the resistance has been replaced by switched capacitor circuit to design switched capacitor integrator as shown in Fig. 3.13

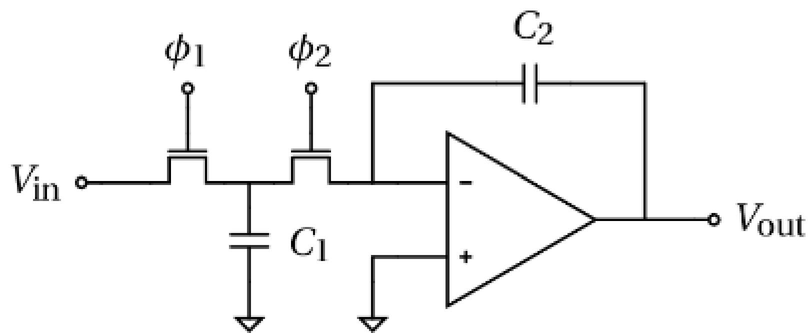


Fig.3.13 : Switched capacitor integrator

Switch capacitor filter

Continuous time 1st order filter is shown in Fig. 3.14

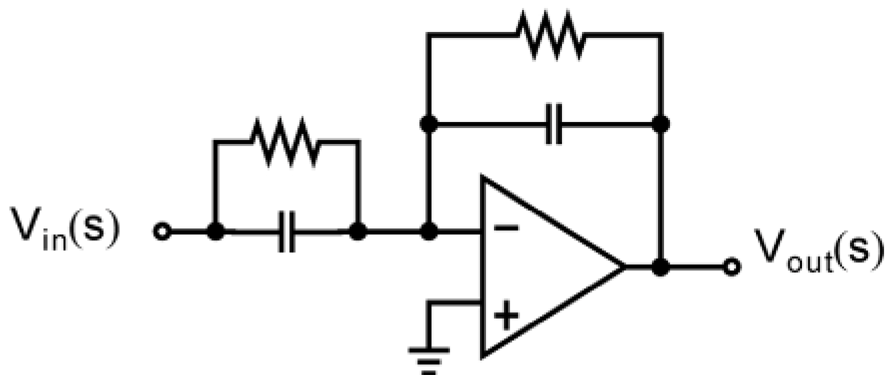


Fig. 3.14 : Continuous time 1st order filter

The switched capacitor of Fig. 3.14 is shown in Fig.3.15

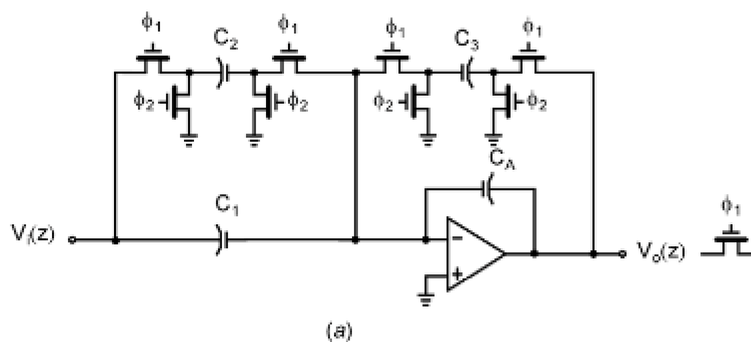


Fig. 3.15 : 1st order switched capacitor filter

3.6. CMOS differential amplifier – design parameters

An amplifier can be characterised with following parameters

Large Signal Voltage Transfer Characteristics

- Maximum Signal Swing Limits
- Small Signal Mid-band Performance Gain
- Input resistance

- Output resistance
- Small Signal Frequency Response
- Other Considerations Noise , Power

The circuit diagram of CMOS differential amplifier is shown in Fig .3.16.

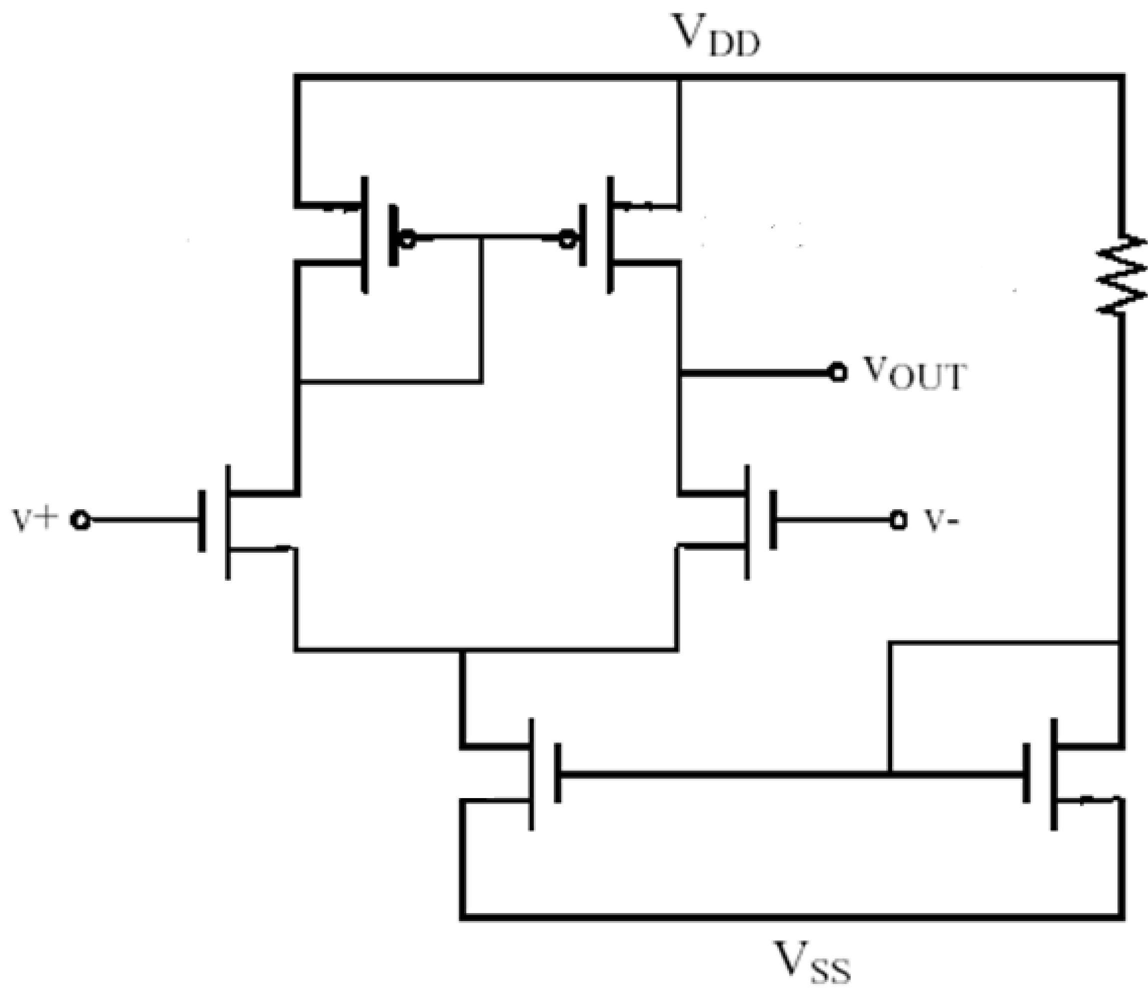


Fig. 3.16 : Circuit diagram of CMOS differential amplifier

The transfer characteristics of differential amplifier is shown in Fig. 3.17

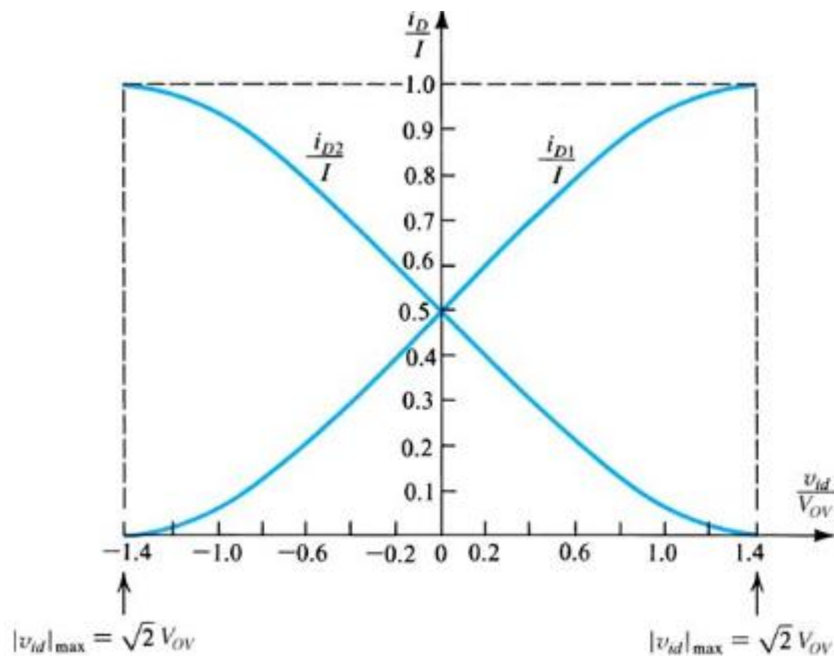


Fig. 3.17 : Transfer characteristics of differential amplifier

3.7 Output amplifier (basic circuit)

Followings are the requirements of output amplifier

- Provide sufficient output power in the form of voltage or current.
- Avoid signal distortion for large signal swings.
- Be efficient.
- Provide protection from abnormal conditions.

Followings are some example of output amplifier

a. Class A amplifier.
b. Source follower.
c. Push-Pull amplifier (inverting and follower)

Circuit diagram of Class-A power amplifier is shown in Fig.3.18

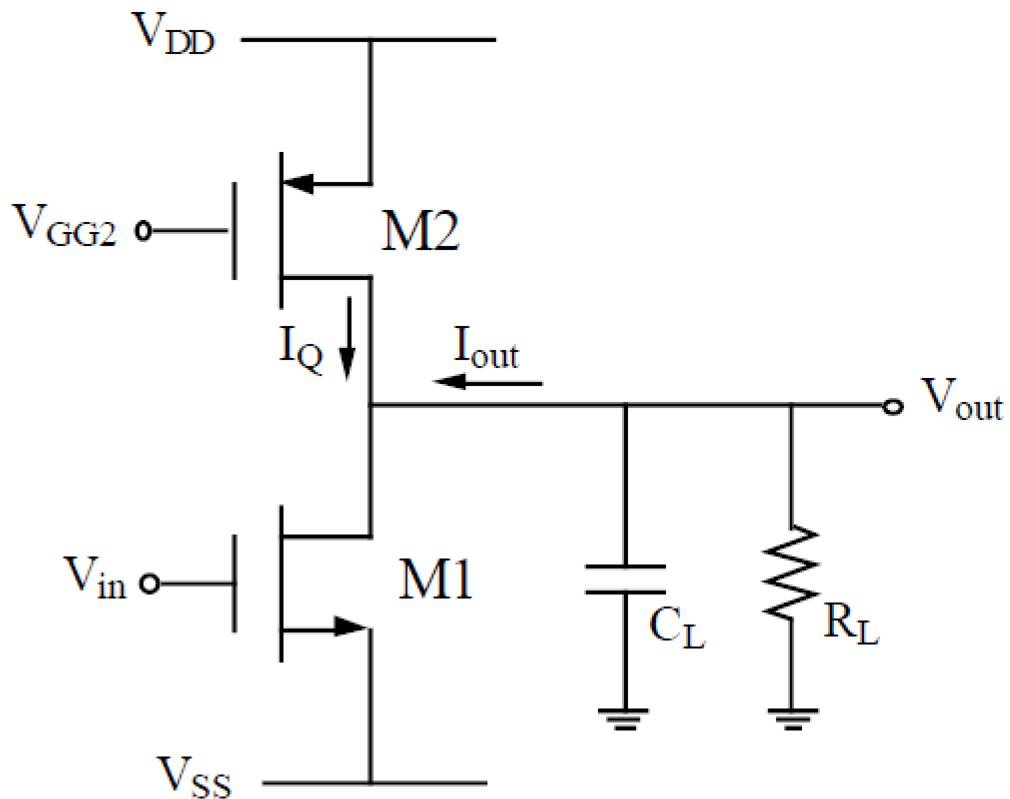


Fig.3.18 : Class –A power amplifier

Efficiency of class A power amplifier is found as

$$\text{Efficiency} = \frac{P_{RL}}{P_{supply}} = \left(\frac{V_{out}(peak)}{(V_{DD} + V_{SS})}\right)^2$$

Circuit diagram of push-pull CMOS amplifier is shown in Fig. 3.19

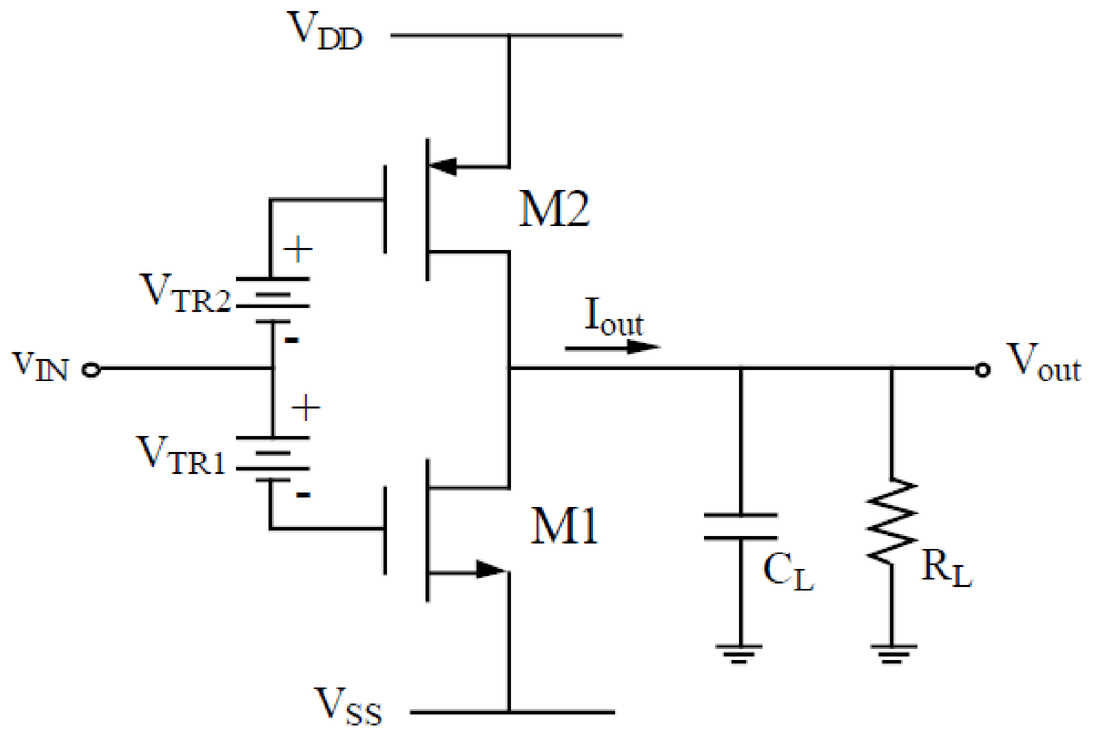


Fig. 3.19 : Circuit diagram of push-pull CMOS amplifier circuit

3.8 Two-Stage CMOS OP-AMP design

Op-amp are designed considering the following specifications

Specifications:

- Gain
- Bandwidth
- Output voltage swing
- PSRR
- Settling time
- CMRR
- Power dissipation
- Noise
- Supply voltage
- Common-mode input range
- Silicon area

The circuit diagram of two-stage op-amp is shown in Fig.3.20

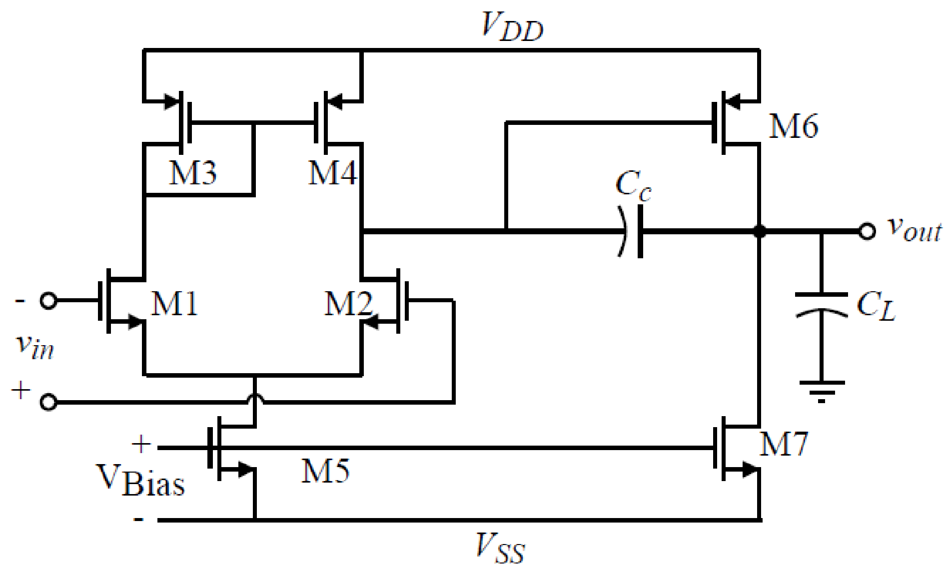


Fig.3.20: Circuit diagram of two-stage op-amp

The design relationship is shown in the circuit diagram of Fig. 3.21

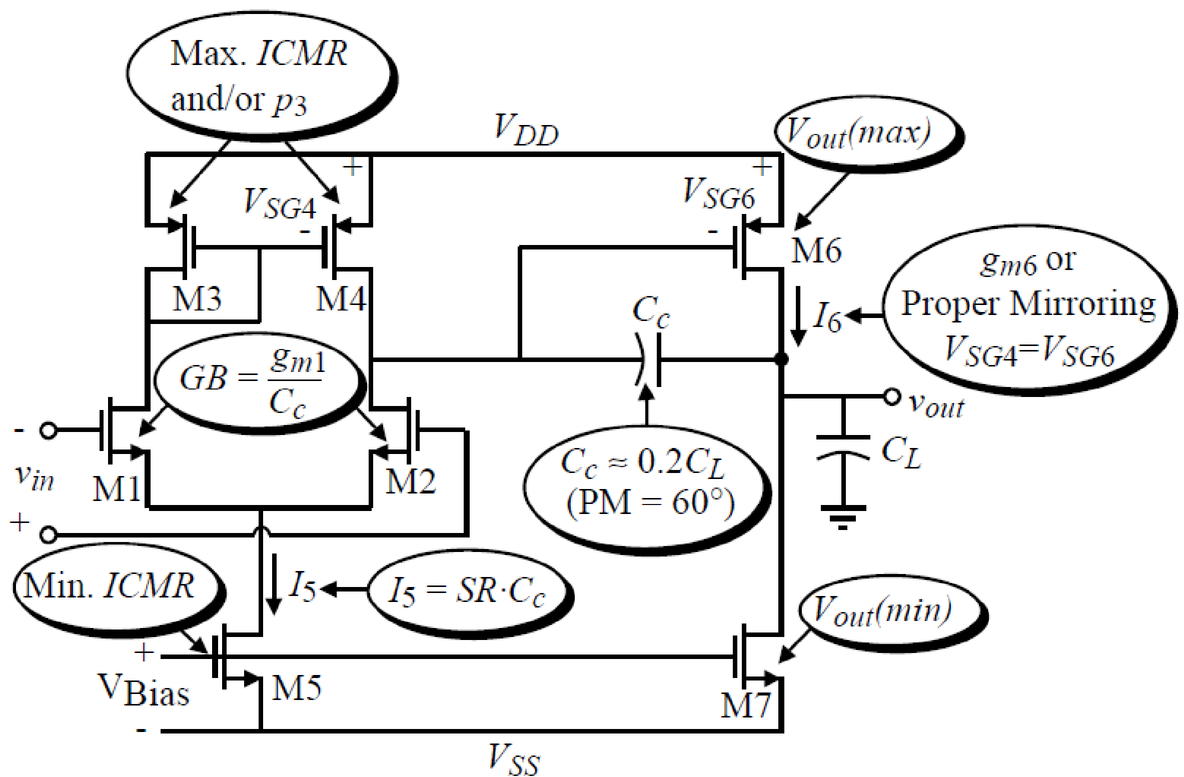


Fig. 3.21: Design relationship is shown in the diagram

Sample Questions (Module -3)

3.1 MCQ

i) MOSFET can be used (a) only as a switch (b)only as an amplifier (c) as both switch and amplifier (d) never for switch nor amplifier

ii) Active resistance means (a) it is a carbon composition resistance (b) it a rheostat made of copper wire (c) it one resistance made with high value material like gold (d) it is realized using MOS transistor

iii) Switch capacitor circuit is used for the emulation of (a) inductor (b) transformer (c) resistance (d) Transistor

iv) Current mirror circuit is cam be configured using

(a) BJT only (b) MOSFET only (c) both BJT and MOSFET (d) neither BJT nor MOSFET

v) CMRR of an Op-Amp ca be calculated considering (a) only common mode gain (b) only difference mode gain (c) Both common mode gain and difference mode gain (d) neither common mode gain nor difference mode gain

3.2 Short answer type questions

i. What is active resistor? Explain the operation of current mirror circuit using MOSFET

ii. How resistance of a constant current source can be increased ?

iii. Write down the requirements of voltage and current references? draw the characteristics of ideal voltage and current references

iv. Explain the operation of band-gap references

v. Draw and explain small signal model of MOSFET

3.3. Ling answer type questions

i. What are the advantages of switch capacitor circuits? How resistance can be emulated using series and parallel switch capacitor circuit ?

ii. Explain the working of switch capacitor integrated circuit. With a neat circuit diagram and wave form explain the working of active loaded CMOS differential amplifier

iii. Discuss about the design parameters of two stage CMOS op-amp

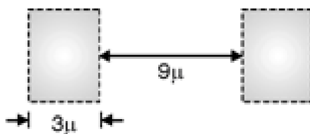
Module -4: Layout Design Rules and Fabrication Steps of ICs

4.1 Micron and Lambda based design rules

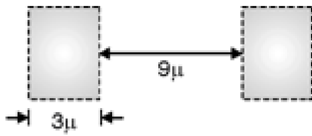
4.1.1 Micron Desin Rules

Micron(μ) Design Rules : Industry uses the micron design rules and code designs in terms of these micron dimensions. The micron design rules are as follows:

(1) Rules for N-well as shown in Figure below.
1. μWidth = 3 μ
2. μSpace = 9 μ

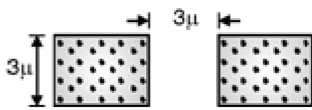


2) Rules for active area as shown in Figure below.
1. μMinimum size = 3 μ
2. μMinimum spacing = 3 μ
2. μN+ active to N-well = 7 μ



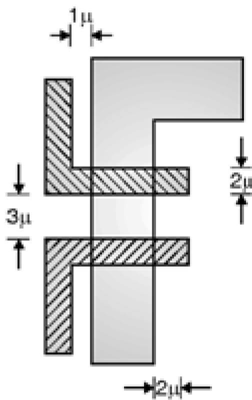
(3) Rules for poly 1 as shown in Figure below.

1. μWidth = 2 μ
2. μSpacing = 3 μ
3. μGate overlap of active = 2 μ
4. μField poly 1 to active = 1 μ



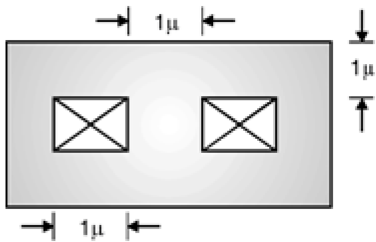
(4) Rules for contact to poly 1 as shown in Figure below.

1. μ 2× μExact contact size = 2 μ
2. μMinimum poly overlap = 1 μ
3. μMinimum contact spacing = 2 μ



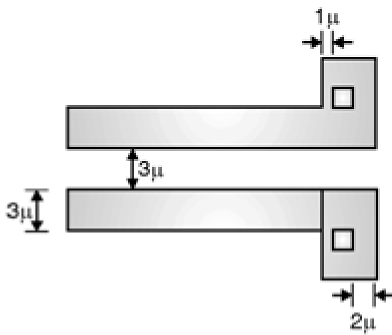
(5) Rules for contact to active as shown in Figure below.

1. μ 2× μExact contact size = 2 μ
2. μMinimum active overlap = 1 μ
3. μMinimum contact spacing = 2 μ
4. μMinimum spacing to gate = 2 μ



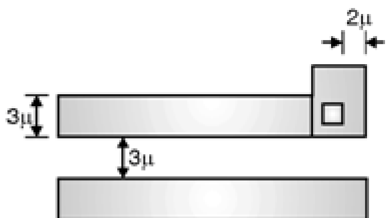
(6) Rules for metal 1 as shown in Figure below.

1. μWidth = 3 μ
2. μSpacing = 3 μ
3. μOverlap of contact = 1 μ
4. μOverlap of via = 2 μ



(7) Rules for metal 2 as shown in Figure below.

1. μWidth = 3 μ
2. μSpace = 3 μ
3. μMetal 2 overlap of via = 2 μ



4.1.2 Lambda Based Design Rules:

In general design rules and layout methodology based on concept of Lambda provide a process and feature size independent way of setting out mask dimension to scale. All parts in all layers will be dimensioned in Lambda(λ) units and subsequently Lambda can be allocated an appropriate value compatible with the feature size of the fabrication process. This concept means that actual mask layout design takes little amount of value subsequently allocated to the feature size, but the design rules are such that it, correctly the mask layouts will provide working circuits for a range of values allocated to Lambda. For example Lambda(λ) can be allocated to a value of 1 Micron so that minimum feature size on cheap will be 2 Micron. Design rules can be conveniently set out in diagrammatic form as shown in figure the widths and separation of conducting paths and for extensions and separation associated with transistor layouts.

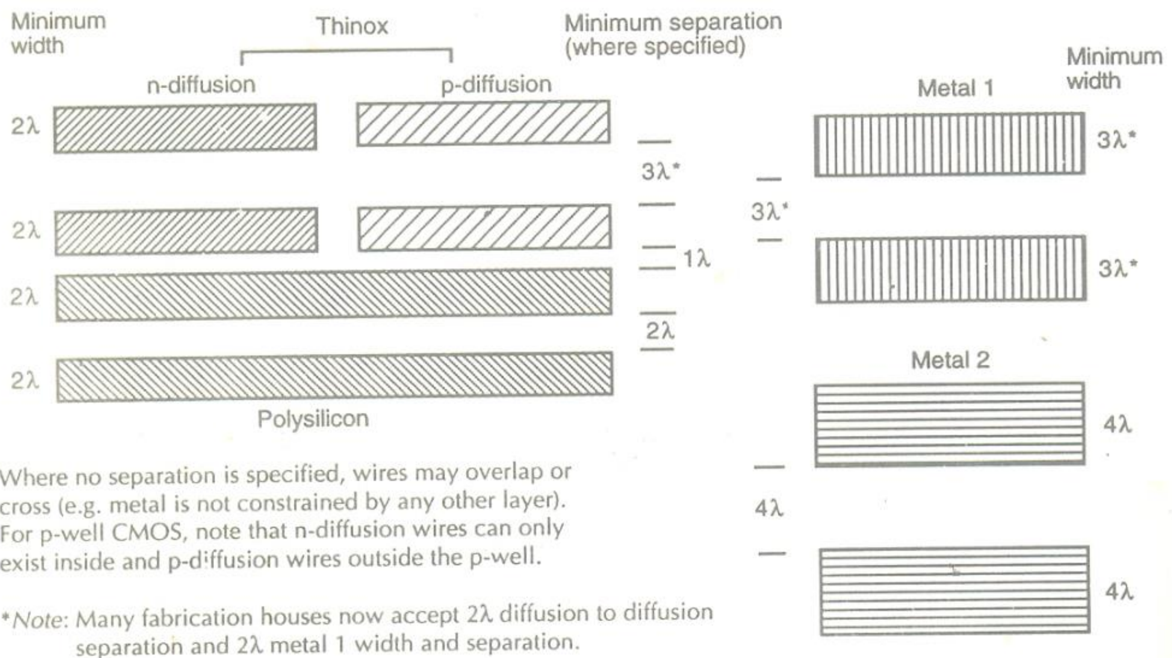


Figure 4.1.2.1: Lambda based design rules for wires

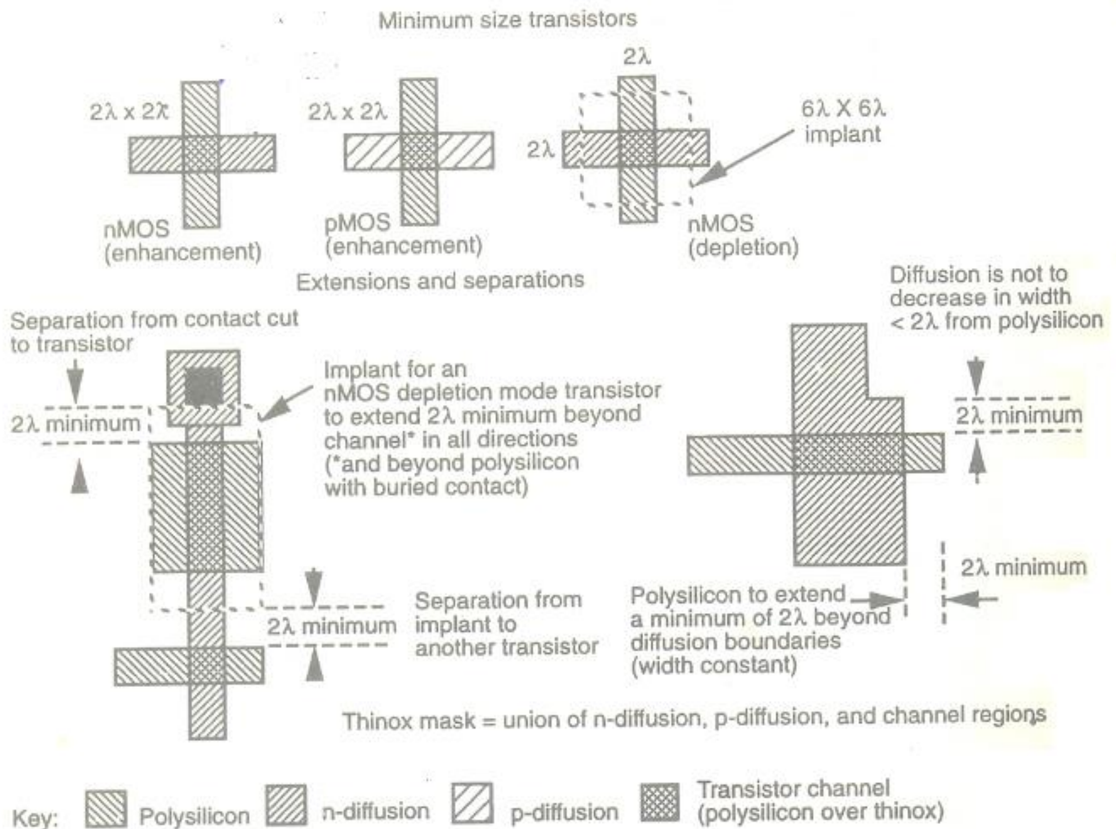
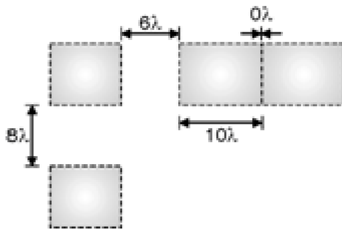


Figure 4.1.2.2 : Design rules for nMOS,pMOS,CMOS

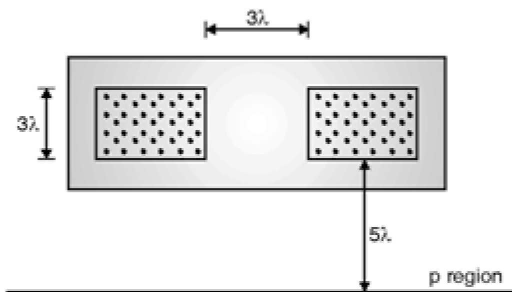
CMOS 'λ' Design Rules : The MOSIS stands for MOS Implementation Service is the IC fabrication service available to universities for layout, simulation, and test the completed designs. The MOSIS rules are scalable λ rules.The MOSIS design rules are as follows :

(1) Rules for N-well as shown in Figure below.

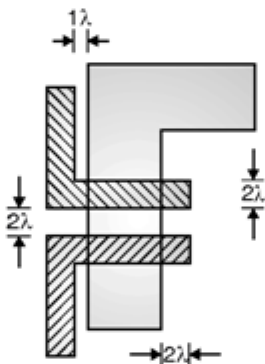
1. Minimum width = 10λ
2. Wells at same potential with spacing = 6λ
3. Wells at same potential = 0λ
4. Wells of different type, spacing = 8λ



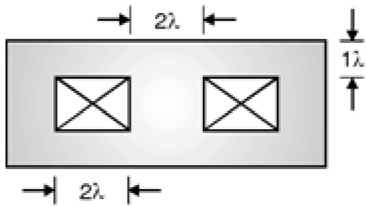
(2) Rules for Active area shown in Figure below.
1. Minimum width = 3λ
2. Minimum spacing = 3λ
3. Source/Drain active to well
edge = 5λ
4. Substrate/well contact active
to well edge = 3λ



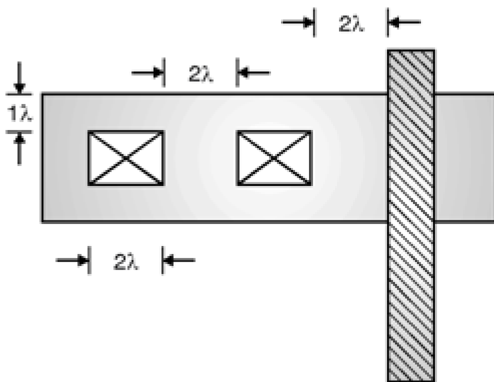
3) Rules for poly 1 as shown in Figure below.
1. Minimum width = 2λ
2. Minimum spacing = 2λ
3. Minimum gate extension of active = 2λ
4. Minimum field poly to active = 1λ



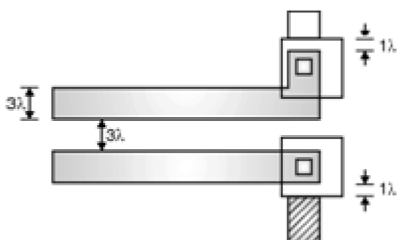
(4) Rules for contact to poly 1 as shown in Figure below.
1. 2 λ×Exact contact size = 2 λ
2. Minimum poly 1 overlap = 1 λ
3. Minimum contact spacing = 2 λ



(5) Rules for contact to active as shown in Figure below. 1. 2λ 2.×Exact contact size = 2λ Minimum active overlap = 1λ 3. Minimum contact spacing = 2λ 4. Minimum spacing to gate of transistor = 2λ

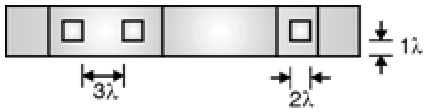


(6) Rules for metal 1 as shown in Figure below.
1. Minimum width = 3λ
2. Minimum spacing = 3λ
3. Minimum overlap of poly contact = 1λ
4. Minimum overlap of active contact = 1λ



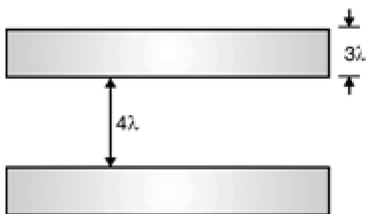
(7) Rules for via 1 as shown in Figure below.

1. λ×Minimum size = 2λ
2. Minimum spacing = 3λ
3. Minimum overlap by metal 1 = 1λ



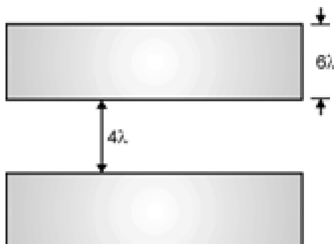
(8) Rules for metal 2 as shown in Figure below.

1. Minimum size = 3λ
2. Minimum spacing = 4λ



(9) Rules for metal 3 as shown in Figure below.

1. Minimum width = 6λ
2. Minimum spacing = 4λ



4.2 Stick diagrams and Layout

Stick diagrams may be used to convey layer information through the use of a color code. For example: n-diffusion--green poly--red blue-- metal yellow--implant black--contact areas.

Encodings for NMOS process shown below:

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

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

Figure 4.2.1: Encoding for single metal nMOS process

Figure 4.2 1 shows the way of representing different layers in stick diagram notation and mask layout using nmos style.Figure 4.2.1 shows when a n-transistor is formed: a transistor is formed when a green line (n+ diffusion) crosses a red line (poly) completely. Figure also shows how a depletion mode transistor is represented in the stick format.

Encodings for CMOS process:

COLOR	STICK ENCODING	LAYERS	MASK LAYOUT ENCODING	CIF LAYER
GREEN	Encoding as in Color plate 1(a)	n-diffusion (n^+ active) Thinox*	* Thinox = n-diff. + p-diff. + transistor channels Encoding as in Color plate 1(a)	CAA or CNA
RED		Polysilicon		CPF
BLUE		Metal 1		CMF
BLACK		Contact cut		CC
GRAY		Overglass		COG
YELLOW (STICK)	green outline here for clarity	p-diffusion (p^+ active)	p^+ mask either or	CAA or CPA
YELLOW	Not shown on diagram	p^+ mask		CPP
DARK BLUE OR PURPLE		Metal 2		CMS
BLACK		VIA		CVA
BROWN	Demarcation line p-well edge is shown as a demarcation line in stick diagrams	p-well		CPW
BLACK	X	V_{DD} or V_{SS} contact	V_{DD} V_{SS}	CC

FEATURE	FEATURE (STICK)	FEATURE (SYMBOL)	FEATURE (MASK)
n-type enhancement mode transistor *(as in Color plate 1(a))*	Demarcation line L: W		
Transistor length to width ratio L:W may be shown.			
p-type enhancement mode transistor	L. W S G D Demarcation line	S G D	S G D p^+ mask
Note: p-type transistors are placed above and n-type below the demarcation line			

Figure 4.2.2: Encoding for double metal CMOS p-well process

There are several layers in an nMOS chip:

_ a p-type substrate

_ paths of n-type diffusion

_ a thin layer of silicon dioxide

_ paths of polycrystalline silicon

_ a thick layer of silicon dioxide

_ paths of metal (usually aluminum)

_ a further thick layer of silicon dioxide

With contact cuts through the silicon dioxide where onnections are required. The three layers carrying paths can be considered as independent conductors that only interact where polysilicon crosses diffusion to form a transistor. These tracks can be drawn as stick diagrams with _ diffusion in green _ polysilicon in red _ metal in blue using black to indicate contacts between layers and yellow to mark regions of implant in the channels of depletion mode transistors.With CMOS there are two types of diffusion: n-type is drawn in green and p-type in brown. These are on the same layers in the chip and must not meet. In fact, the method of fabrication required that they be kept relatively far apart. Modern CMOS processes usually support more than one layer of metal. Two are common and three or more are often available. Actually, these conventions for colors are not universal; in particular, industrial (rather than academic) systems tend to use red for diffusion and green for polysilicon. Moreover, a shortage of colored pens normally means that both types of diffusion in CMOS are colored green and the polarity indicated by drawing a circle round p-type transistors or simply inferred from the context. Colorings for multiple layers of metal are even less standard. There are three ways that an nMOS inverter might be drawn:

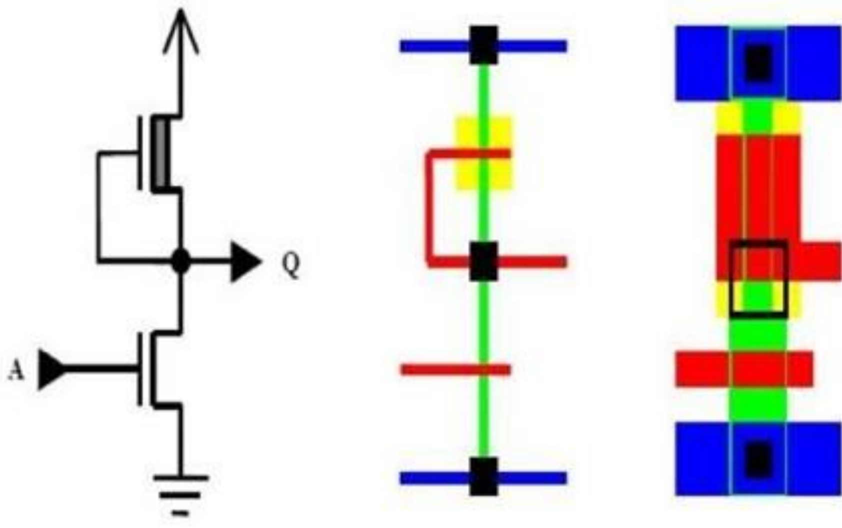


Figure 4.2.3: nMOS depletion load inverter.

Figure4 shows schematic, stick diagram and corresponding layout of nMOS depletion load inverter.

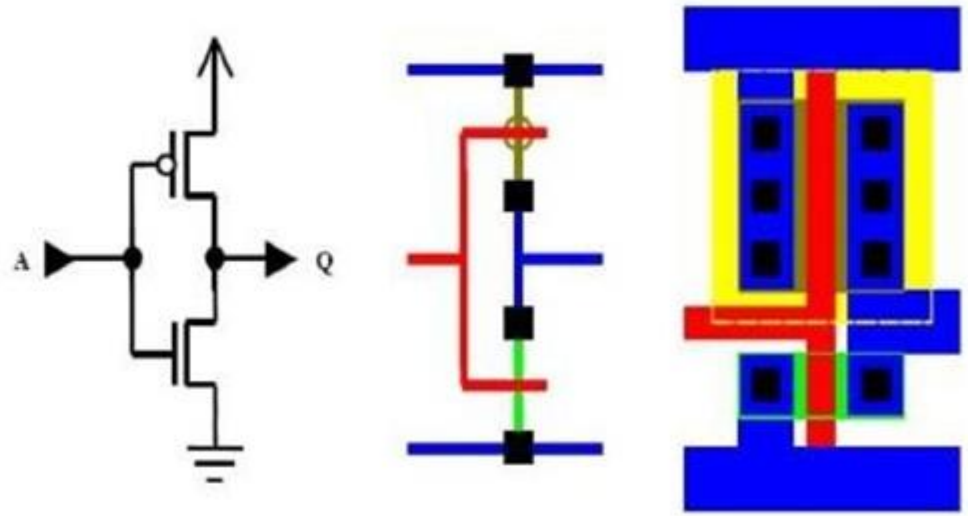


Figure 4.2.4: CMOS inverter

Figure 4.2.4 shows the schematic, stick diagram and corresponding layout of CMOS inverter.

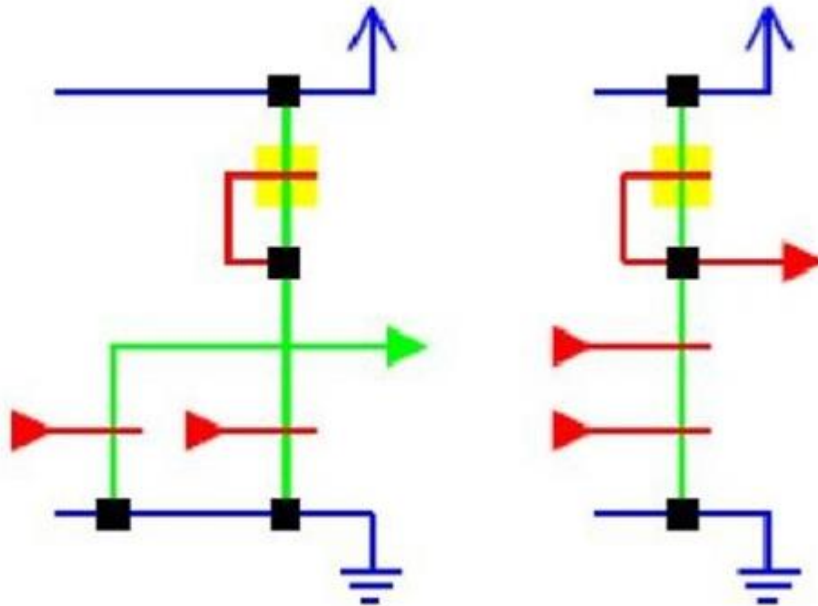


Figure 4.2.5: The stick diagrams for nMOS NOR and NAND.

Layout

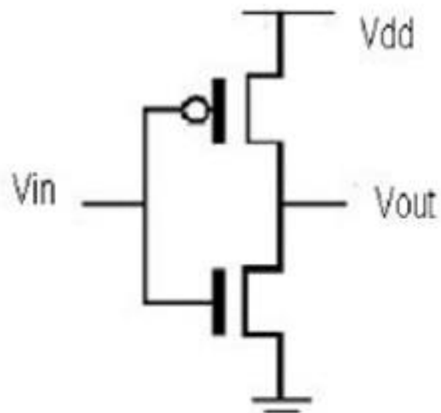


Figure:4.2.6 CMOS inverter schematic

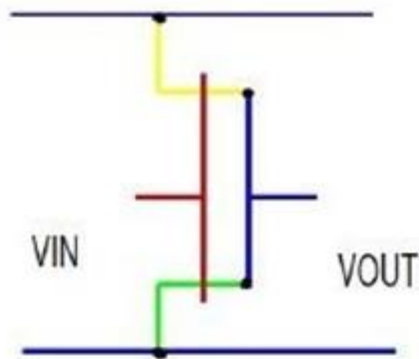


Figure:4.2.7 CMOS inverter stick diagram

The figure 4.2.7 shown here is the stick diagram for the CMOS inverter. It consists of a Pmos and a Nmos connected to get the inverted output. When the input is low, Pmos (yellow) is on and pulls the output to vdd; hence it is called pull up device. When Vin =1, Nmos (green) is on it pulls Vout to Vss, hence Nmos is a pull down device. The red lines are the poly silicon lines connecting the gates and the blue lines are the metal lines for VDD (up) and VSS (down).The layout of the cmos inverter is shown below. Layout also gives the minimum dimensions of different layers, along with the logical connections and main thing about layouts is that can be simulated and checked for errors which cannot be done with only stick diagrams.

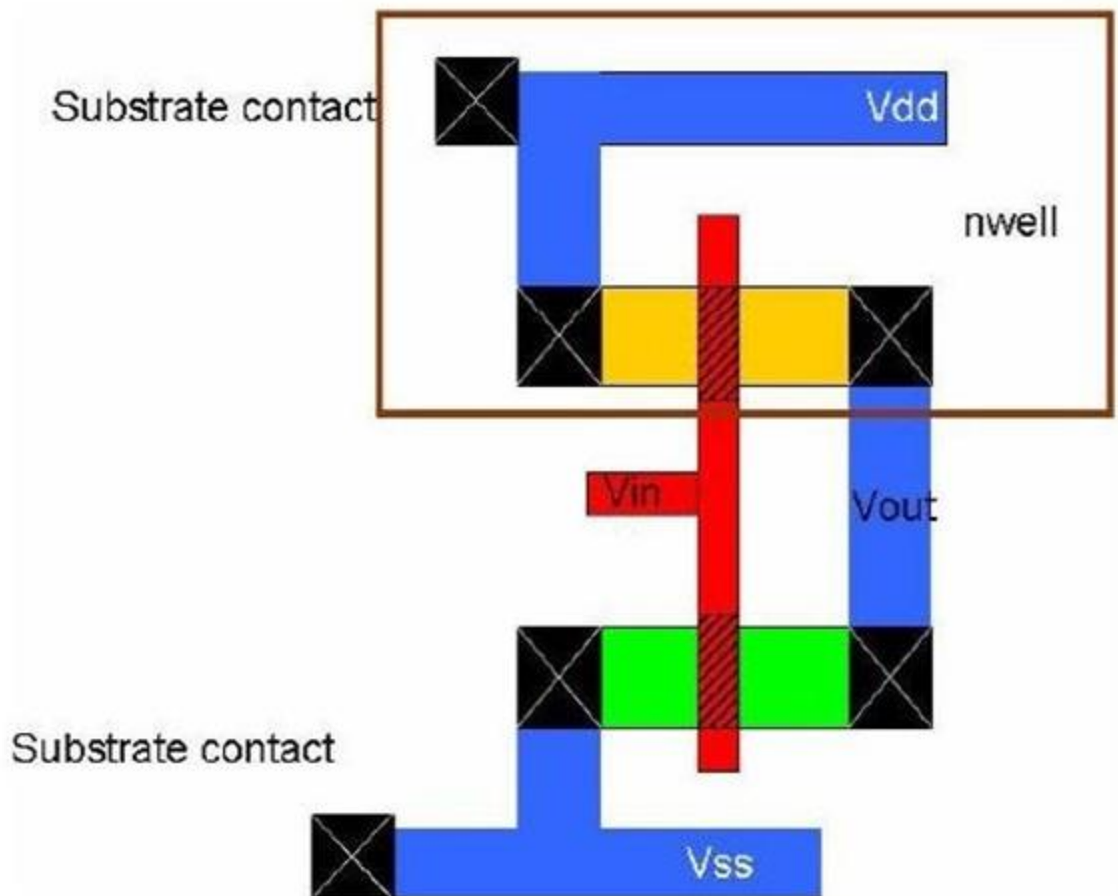


Figure 4.2.8:Layout of CMOS Inverter

The layout shown above is that of a CMOS inverter. It consists of a pdiff (yellow colour) forming the pmos at the junction of the diffusion and the polysilicon (red colour) shown hatched ndiff (green) forming the nmos(area hatched).The different layers drawn are checked for their dimensions using the DRC rule check of the tool used for drawing. Only after the DRC (design rule check) is passed the design can proceed further. Further the design undergoes Layout Vs Schematic checks and finally the parasitic can be extracted. Figure 22: Schematic diagrams of nand and nor gateWe can see that the nand gate consists of two pmos in parallel which forms the pull up logic and two nmos in series forming the pull down logic. It is the complementary for the nor gate. We get inverted logic from CMOS structures. The series and parallel connections are for getting the right logic output. The pull up and the pull down devices must be placed to get high
And outputs when required. Figure 4.2.9: Stick diagrams of nand gate.

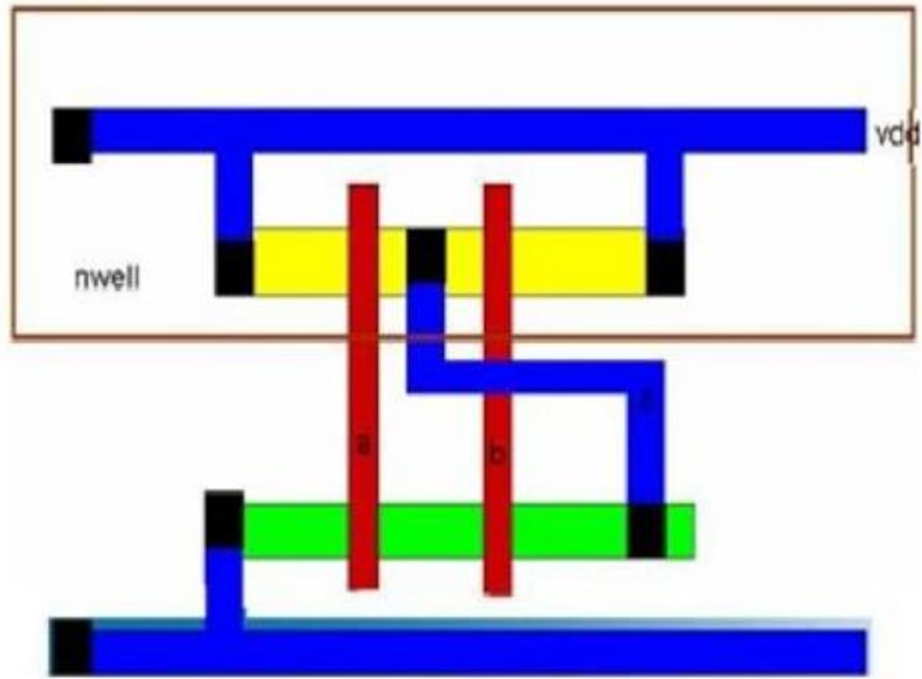


Figure 4.2.10: Layout of nand gate.

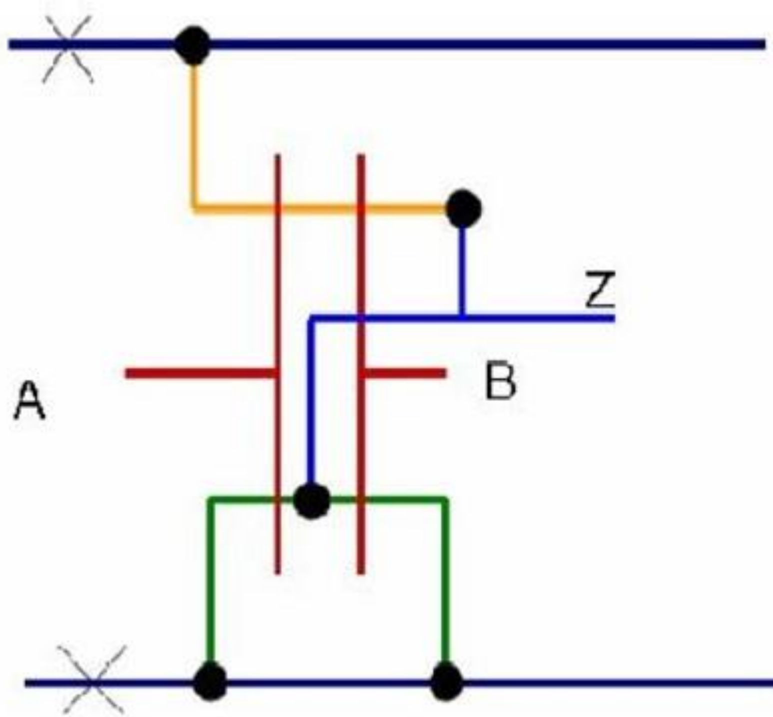


Figure 4.2.11: Stick diagram of nor gate.

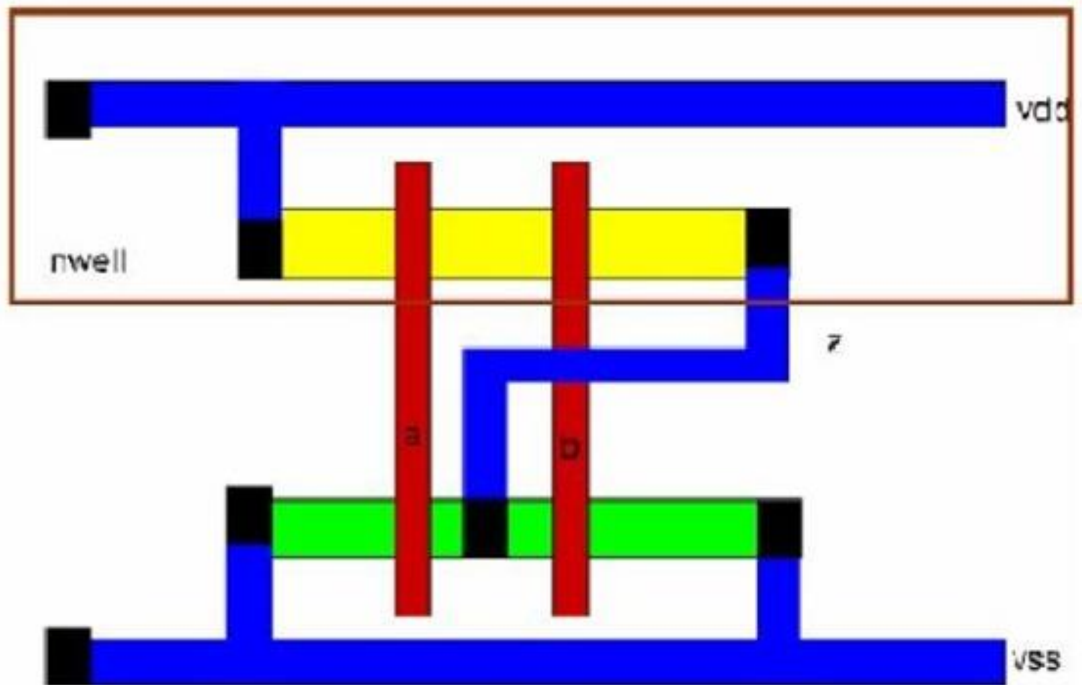


Figure 4.2.12: Layout of nor gate

4.3 Fabrications steps of IC

4.3.1 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.

4.3.1.1 Czochralski technique(CZ)-manufacturing single crystals. It is especially suited for the largewafers that are currently used in ICfabrication.

4.3.1.2 Float zone technique - this is mainly used for small sized wafers. The float zone technique is used for producing specialty wafers that have low oxygen impurity concentration.

4.3.1.1 Czochralski crystal growth technique

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 SiO_2or SiN_x. The drawback is that at the high temperature the inner liner of the crucible also starts melting and has to replaced periodically.

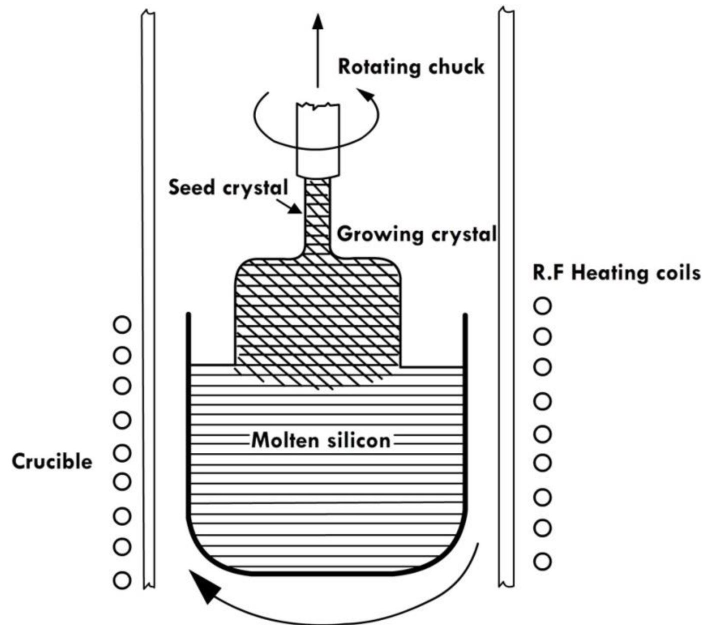


Figure 4.3.1.1 Schematic of the Czochralski growth technique.

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.

Figure 4.3.1.2: Single crystal Si ingot.

furnace is heated above 1500 ˚C, since Si melting point is 1412 ˚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.

4.3.1.2 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 figure 4.3.1.3 .Apolycrystalline 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 4.3.1.3.Whenthe 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.

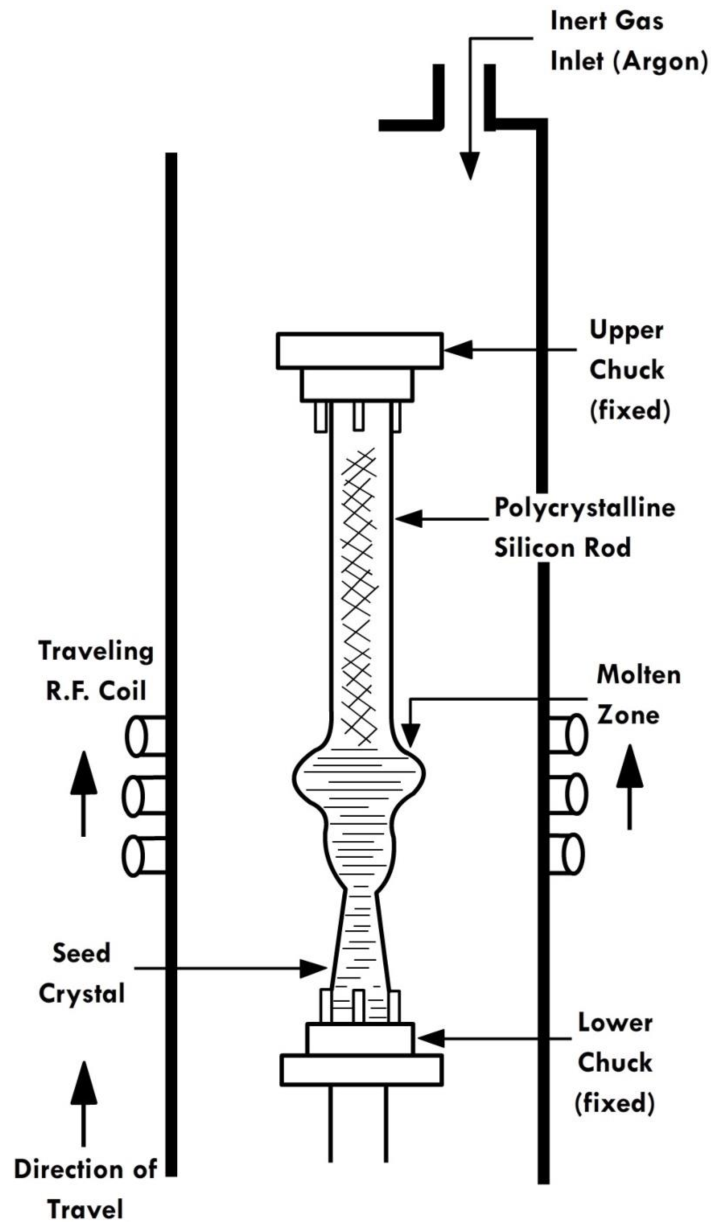


Figure 4.3.1.3: Schematic of the float zone technique.

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.

4.3.1.3 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 an 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 4.3.1.4.*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 4.3.1.4. Larger wafers are usually thicker, for mechanical integrity. After cutting, the wafers are chemically etched to remove any damaged and

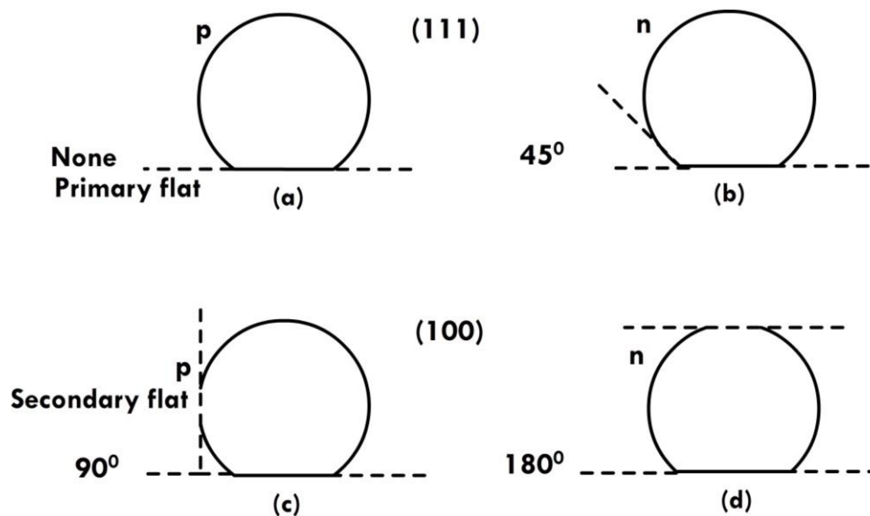


Figure 4.3.1.4:Flats for the different wafer types and orientations.All orientations and doping types have a primary flat,while there are different secondary flats for different types (a) p(111) (b) n(111) (c) p(100) and (d) n(100).

Table 4.3.1: Specs of a typical 150 mm wafer

Specs	Value
Diameter	150 ± 0.5 mm
Thickness	$675 \pm 25\ \mu m$
Orientation	$100 \pm 1^{\circ}$
Bow	$60\ \mu m$
Δt	$50\ \mu m$
Primary flat	55-60 mm

Secondary flat	35-40 *mm*

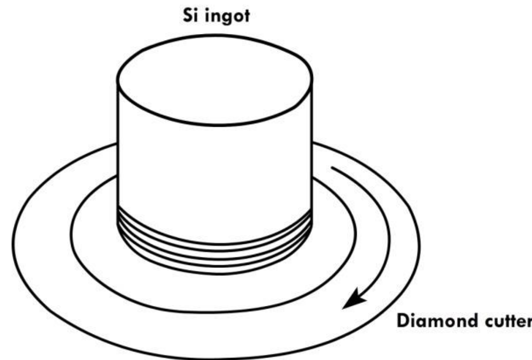


Figure 4.3.1.5: Inner diameter wafer slicing, used for cutting the ingots into individual wafers.

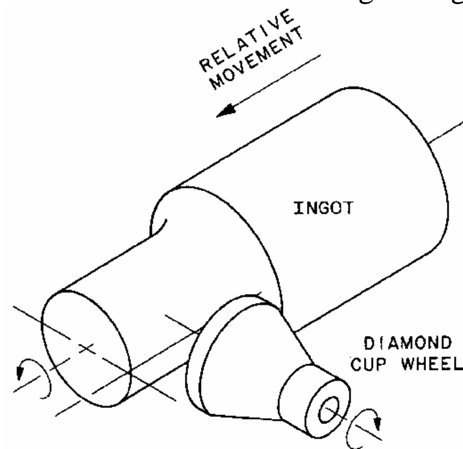
contaminated regions. This is usually done in an acid bath with a mixture of hydrofluoric acid, nitric acid, and acetic acid. After etching, the surfaces are polished, first a rough abrasive polish, followed by a chemical mechanical polishing (CMP) procedure. In CMP, a slurry of fine SiO_2particles suspended 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.

4.3.1.4 Wafer Preparation

Silicon, albeit brittle, is a hard material. The most suitable material for shaping and cutting silicon is industrial-grade diamond. Conversion of silicon ingots into polished wafers requires several machining, chemical, and polishing operations. After a grinding

operation to fix the diameter of the material (***Figure 4.3.1.6***), one or more flats are grounded along the length of the ingot. The largest flat, called the "major" or "primary" flat, is usually relative to a specific crystal orientation. The flat is located by x-ray diffraction techniques. The primary flat serves as a mechanical locator in automated processing equipment to position the wafer, and also serves to orient the IC device relative to the crystal. Other smaller flats are called "secondary" flats that serve to identify the orientation and conductivity type of the wafer. Secondary flats thus provide a means to quickly sort and identify wafers should mixing occur. The flat locations for the four common types of silicon wafers are exhibited in ***Figure 4.3.1.7***

Figure 4.3.1.6: Schematic of the grinding process.



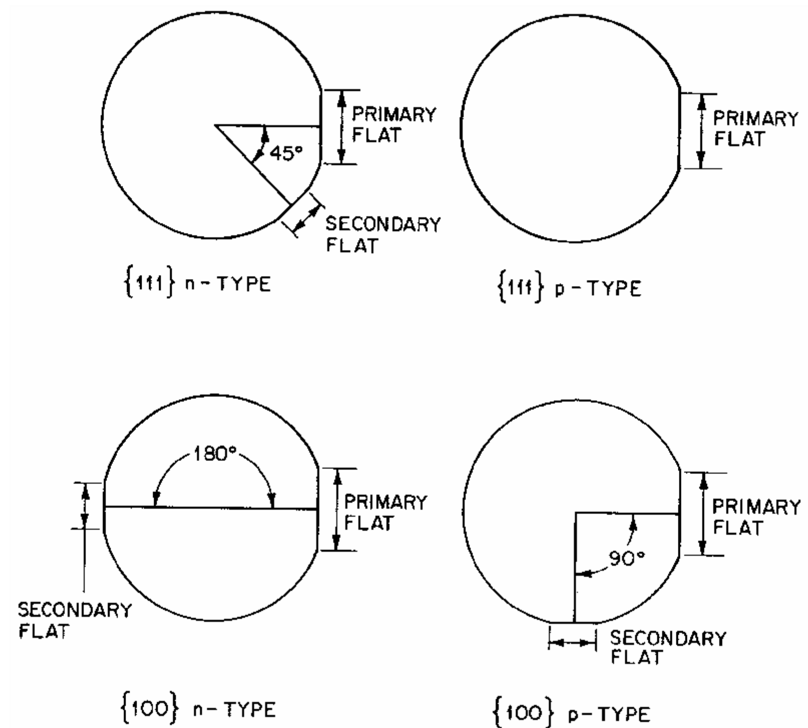


Figure 4.3.1.7: Identifying flats on silicon wafers.

The drawback of these flats is that the usable area on the wafer, i.e. the area on which microelectronic devices can be fabricated, is reduced. For some 200 mm and 300 mm diameter wafers, only a small notch is cut from the wafer to enable lithographic alignment but no dopant type or crystal orientation information is conveyed.Once these operations have been completed, the ingot is ready to be sliced by diamond saw into wafers. Slicing determines four wafer parameters: surface orientation (e.g., <111>or <100>); thickness (e.g., 0.5 – 0.7 mm, depending on wafer diameter); taper, which is the wafer thickness variations from one end to anther; and bow, which is the surface curvature of the wafer measured from the center of the wafer to its edge. After slicing, the wafers undergo lapping operation that is performed under pressure using a mixture of Al_2O_3 and glycerine. Subsequent chemical etching removes any remaining damaged and contaminated regions. Historically, mixtures of hydrofluoric, nitric, and acetic acids have been employed, but alkaline etching, using potassium or sodium hydroxide, is also common.Polishing is the final step.Its purpose is to provide a smooth, specular surface on which device features can be photoengraved. ***Figure 4.3.1.8*** depicts the schematic of a typical polishing machine and the process. ***Figure 4.3.1.9*** displays finished silicon. Wafers of various dimensions. A finished wafer is subject to a myriad of physical tolerances, and examples of the specifications are shown in ***Table 4.2***.

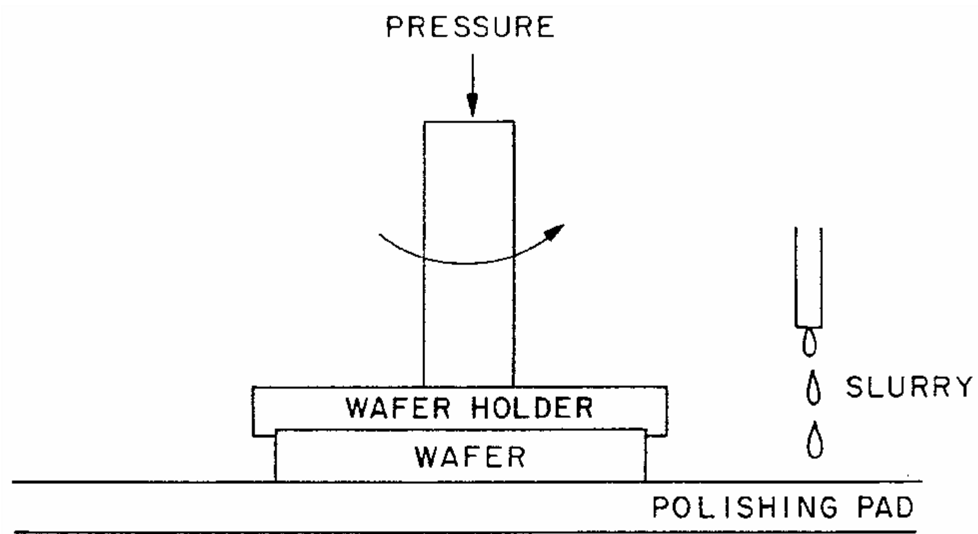


Figure 4.3.1.8: Schematic of the polishing process.

Figure 4.3.1.9: Finished silicon wafers of various sizes.

Table 2: Typical specifications for mono crystalline silicon wafers.

Parameter	**125 mm**	**150 mm**	**200 mm**	**300 mm**
Diameter (mm)	125±1	150±1	200±1	300±1
Thickness (mm)	0.6-0.65	0.65-0.7	0.715-0.735	0.755-0.775
Bow (µm)	70	60	30	<30
Total thickness variation (µm)	65	50	10	<10
Surface orientation	±1°	±1°	±1°	±1°

4.3.2 Oxidation

4.3.2.1 Introduction

Oxidation refers to the conversion of the silicon wafer to silicon oxide (SiO_2 or more generally SiO_x). The ability of Si to form an oxide layer is very important since this is one of the reasons for choosing Si over Ge. The Horni transistor design, which was used in the first integrated circuit by Robert Noyce, was made of Si and the formation of SiO_x helped in fabricating a planar device.
Si exposed to ambient conditions has a *native oxide* on its surface. The native oxide is approximately 3 *nm* thick at room temperature. But this is too thin for most applications and hence a thicker oxide needs to be grown. This is done by consuming the underlying Si to form SiO_x. This is a *grown layer*. It is also possible to grow SiO_x by a chemical vapor deposition process using Si and O precursor molecules. In this case, the underlying Si in the wafer is not consumed. This is called a *deposited layer*.SiO_x helps in protecting the wafer from contamination, both physical and chemical. Thus, it acts as a *passivating layer*. The oxide layer protects the wafer surface from scratches and it also prevents dust from interacting with the wafer surface, and thus minimizes contamination. The oxide layer also protects the wafer from chemical impurities, mainly electrically active

Table 1: Silicon oxide thickness chart

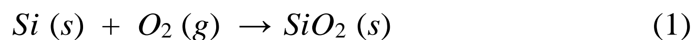
Thickness, in Å	Application
60-100	Tunneling gates
150-500	Gate oxides, capacitor dielectrics
200-500	LOCOS pad oxide
2000-5000	Hard masks, surface passivation
3000-10000	Field oxides

contaminants. SiO_x acts as a hard mask for doping and as an etch stop during patterning. The original gate oxide in MOSFET was made of SiO_x. SiO_x was also used as the inter-layer dielectric separating different metallization layers, though this is usually a deposited layer. SiO_2 is also used to prevent induced charge due to the metal layers, this is called a *field oxide*. In all of these forms, different thickness of the oxide layer are required. These are summarized in table 1.

4.3.2.2 Oxidation types

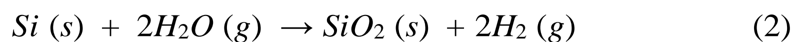
In the case of grown oxide layers, there are two main growth mechanisms

1. **Dry oxidation** – Si reacts with O_2 to form SiO_2.



$$Si\ (s)\ +\ O_2\ (g)\ \rightarrow SiO_2\ (s) \quad (1)$$

2. **Wet oxidation** – Si reacts with water (steam) to form SiO_2.



$$Si\ (s)\ +\ 2H_2O\ (g)\ \rightarrow SiO_2\ (s)\ +\ 2H_2\ (g) \quad (2)$$

In both cases, Si is supplied by the underlying wafer. Dry and wet oxidation need high temperature (900 - 1200 °C) for growth, though the kinetics are different, which is why this process is called **thermal oxidation**. Since the underlying Si is consumed, the Si/SiO_2 interface moves deeper into the wafer. The movement of the interface is shown in 4.3.11.

There is also a volume expansion since the densities of the oxide layer and silicon are different. Thus, the final thickness is higher than the initial Si thickness. Consider the oxide layer silicon interface as shown in figure 4.3.2.2 Here, d is the thickness of the original Si layer that has been consumed in forming the oxide layer of thickness d^{j}. Si has a density of 2.33 gcm^{-3} (ρ_{Si}) and an atomic weight of 28.08 $gmol^{-1}$ (Z_{Si}) while SiO_2 has a density of 2.65 gcm^{-3} (ρ_{SiO_2}) and a molecular weight of 60.08 $gmol^{-1}$ (Z_{SiO_2}). Given that

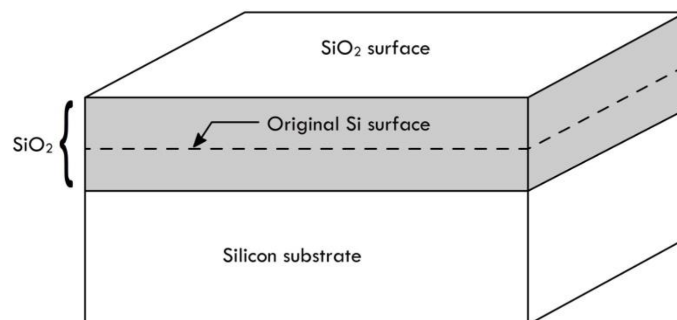


Figure 4.3.2.1: Movement of the silicon-oxide interface as oxide thickness grows.

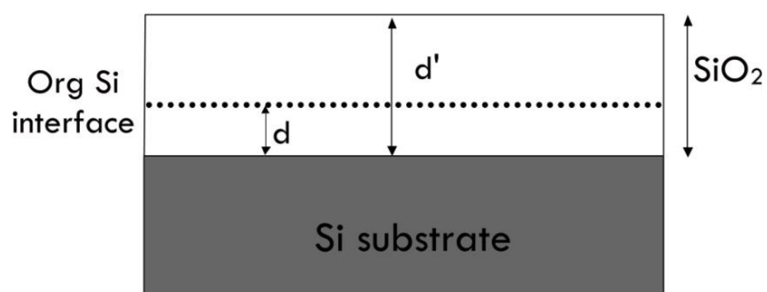


Figure 4.3.2.2: Schematic cross section of the Si oxide interface

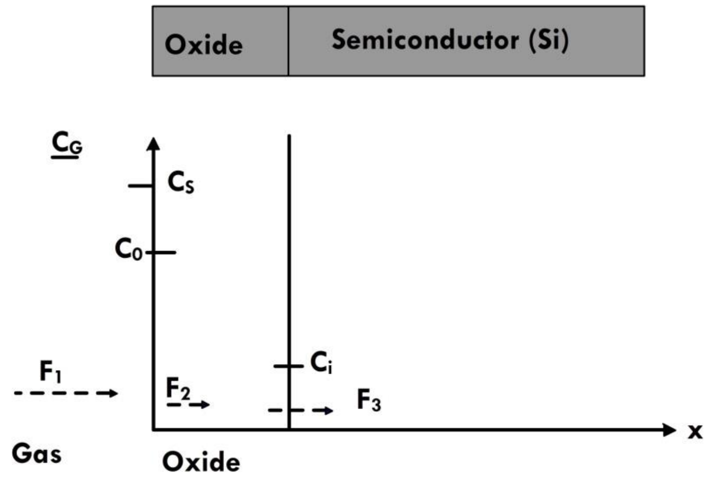


Figure 4.3.2.3: A one dimensional growth model for oxide formation with the fluxes and concentrations marked.

the cross section area, *A*, is the same it is possible to use the law of molar conservancy to derive the relation between *d* and d'. This is given by

$$\frac{dA\rho_{Si}}{Z_{Si}} = \frac{d'A\rho_{SiO2}}{Z_{SiO_2}} \tag{3}$$

Substituting the values in equation 3 gives the relation

$$d' = 1.88\, d \tag{4}$$

Hence, the thickness of the oxide layer is larger than the thickness of the Si that is consumed to form that oxide. To grow 100 *nm* of oxide, per equation 3, 53.2 *nm* of Si needs to be consumed.

4.3.2.3 Oxide furnaces

Thermal oxides are grown in *tube furnaces* by a *batch process*, i.e. multiple wafers are processed at the same time. This becomes important in the con- text of process control, since any deviation from required conditions would affect multiple wafers and hence lead to overall cost increase. For small wafer sizes, typically 3" and 4" wafers, horizontal tube furnaces are used for oxidation. The furnace is typically divided into 3 zones - *source zone, center zone, and load zone*. The source zone is used for introducing the gases required for oxida- tion. Typically, this is oxygen (dry ox) or steam (wet ox) at the appropriate partial pressure (concentration). Sometimes, chlorinated oxide layers are also grown. The chlorine incorporated in the oxygen reduces mobile ions in the oxide layer and also reduces charge concentration at the oxide-Si interface.

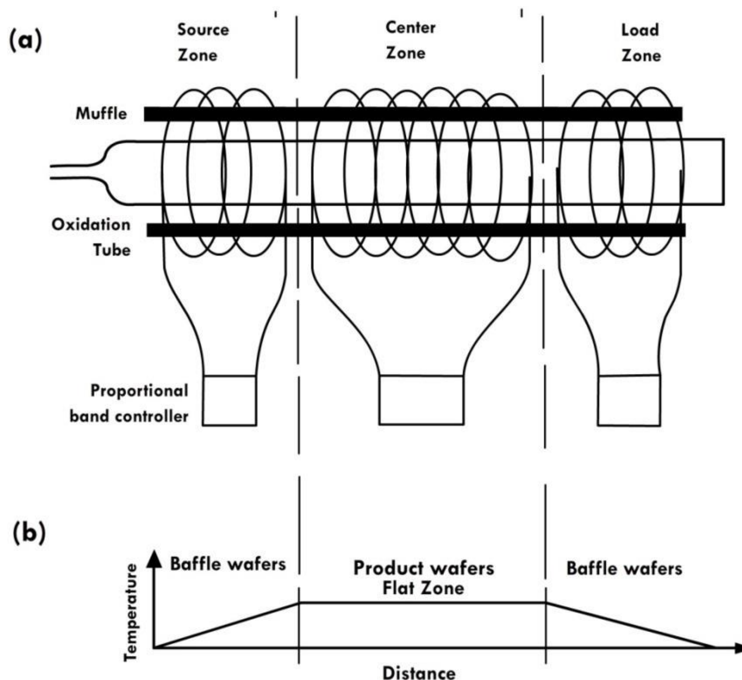


Figure 4.3.2.4: Schematic of (a) horizontal diffusion furnace. (b) The furnace is typically divided into 3 zones, with the product wafers loaded in the center zone. The zones have their individual heaters and temperature controllers, to ensure uniform temperature.

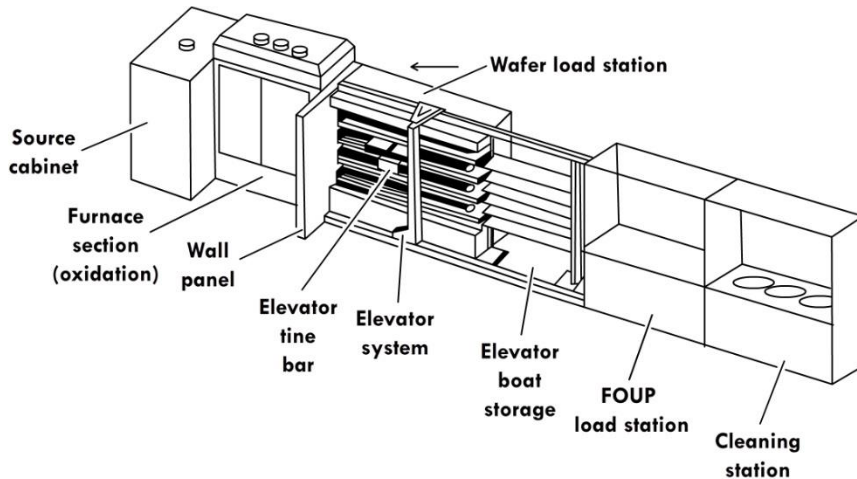


Figure 4.3.2.5: Schematic of a commercial horizontal diffusion furnace. The wafers are loaded in plastic boxes called FOUPs (Front opening universal pod). The wafers are cleaned first before loading in the furnace. In this setup, there is no provision for storing the wafers but newer designs based on vertical furnaces also have provision for storing multiple FOUPs.

This improves cleanliness and device performance. Chlorine is introduced in the form of Cl_2, hydrogen chloride gas (HCl), trichloroethylene (lq), or trichloroethane (lq). Vapor from gaseous sources is mixed with the oxygen, while for liquid sources, the gas is bubbled through the liquid. There are a few purge and pump steps, to reduce contamination in the furnace before oxygen is introduced. Commercial tube furnaces also have loading zones (for loading wafers) and cleaning stations, and stations for storing wafers. A schematic of a commercial horizontal furnace is shown in figure 4.3.2.5.

The process wafers (wafers that are used to fabricate the integrated cir- cuits) are loaded in the center zone. Usually, *baffle plates* are loaded at the ends (these are made of quartz). Bare wafers, called *fillers*, are also loaded along with the process wafers. These help in regulating gas flow through the furnace so that oxide growth is uniform in the process wafers. Thus, not all wafers in the furnace are process wafers. Higher the ratio of process wafers to blank wafers that can be loaded

in the furnace, higher is the *process through- put* (number of process wafers processed per hour). Temperatures are also constantly maintained and regulated within the furnace during oxidation, using a PID (proportional-integral-derivative) mechanism. Typical temperature profile during oxidation is shown in figure 4.3.2.6 The idle temperature is

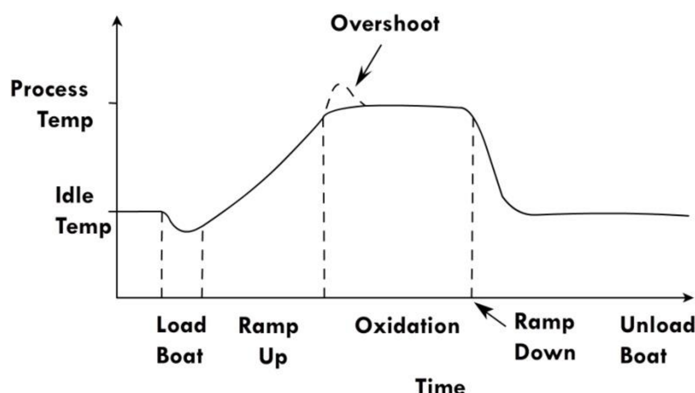


Figure 4.3.2.6: Temperature profile in a tube furnace during oxidation.

typically 300-400 ˚C, to minimize heating and cooling time during the oxida- tion process. The elevated temperature also prevents gases from condensing on the tube inner walls.

For larger wafers (typical process wafers are now 12" or 300 *mm* wafers), hor- izontal furnaces are not practical and also occupy a lot of space. Diffusion furnaces designs are now vertical, called *vertical diffusion furnaces* (VDF). A schematic of a VDF is shown in figure 4.3.2.7 The furnace consists of a loading station and space for storing wafers (before and after processing). The boat (where wafers are loaded for processing) moves vertically into the furnace, see figure 4.3.2.7 VDF are more compact than horizontal furnaces. Gas flow is also more uniform, with less turbulence. The boat is rotated during operation to ensure uniformity of the oxide layer. This is especially true for mixed gases since the gases move parallel to gravity and hence do not get separated. The operation of the VDF is similar to the horizontal tube furnace. There are also baffles and blanks, loaded along with the process wafers. Typically, a 125 wafer boat holds a maximum of 75 product wafers, the rest are fillers, baffles, and monitor wafers (for measuring oxide thickness and uniformity for process control).

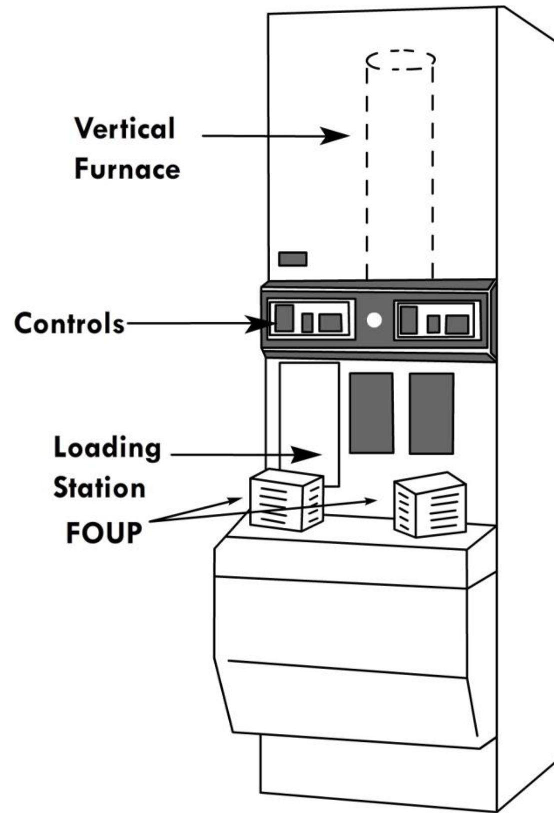


Figure 4.3.2.7: Schematic of a vertical diffusion furnace.

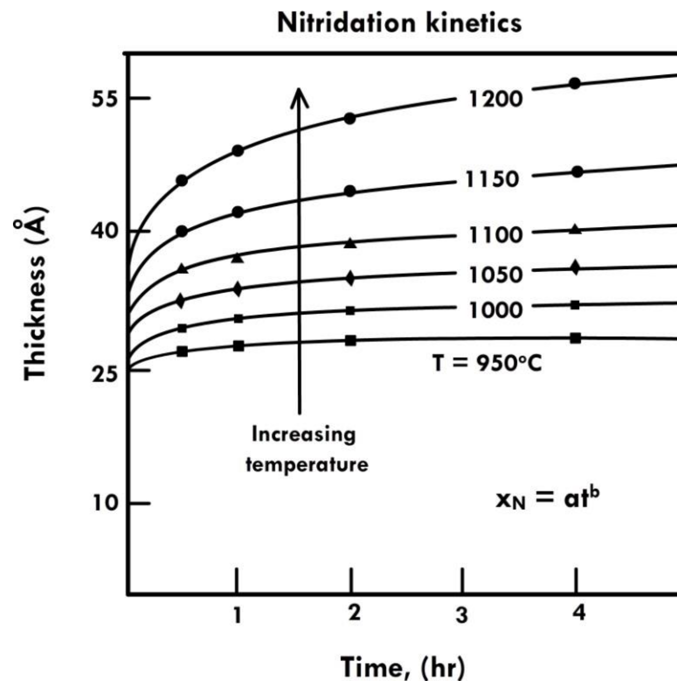


Figure 4.3.2.8: Nitride thickness vs time for different temperatures. The rate is smaller than oxidation since the gas has to diffuse through the nitride layer, and that forms the rate limiting step.

4.3.3 Photolithography

4.3.3.1 Introduction

Lithography (or patterning) refers to the *series of steps* that establish the *shapes, dimensions, and location* of the various components of the integrated circuit (IC). The current progress in IC design, with the decreased dimensions (miniaturization) of the chip and increased density of transistors, is possible only if smaller areas on the wafer surface can be patterned. This is primarily the function of lithography. Thus, the success of modern IC design is due largely to lithography. This can be summarized in the process goals

1. Create a pattern with the dimensions established by the circuit design.
2. Place the pattern correctly with respect to the crystal orientation and other existing patterns.

After the pattern is created, either the defined part of the wafer surface is removed (trench creation) or left behind (island creation) or new material is deposited. Lithography is also used to expose certain parts of the wafer surface for doping (either with a hard mark for thermal diffusion or with a soft mask for ion implantation).
The correct placement of the circuit pattern involves alignment or *registration* of various masks. An IC wafer fabrication process can require forty or more patterning steps. Alignment of these individual steps is critical to form a working IC.

4.3.3.2 Process overview

For lithography processing, a hard copy of the pattern has to be first gen- erated. This is called a *reticle* or *mask*. The design on the mask has to be transferred to the wafer, as shown in figure 4.3.3.1 The transfer can be 1:1 (i.e. with no reduction in size) but usually the size is reduced so that the pattern is transferred to a smaller region on the wafer. This is done by using suitable lens to demagnify the pattern. Lithography can be broadly divided into two stages, each of which consists of several steps.

2. First, the pattern is transferred to a *photoresist layer* on the wafer. Photoresist is a light sensitive material whose properties change on exposure to light of specified wavelength. This process is called *de- veloping* . The pattern formed in this step is temporary and can be removed easily. This is especially important if the pattern is not prop- erly alignment with the wafer or with any existing patterns on the wafer, *improper registry*.

3. The transfer of the pattern takes place from the photoresist to the wafer. Exposed wafer surfaces can be etched (removal of material) or layers deposited on it. Dopant materials can also be added to sections of the wafer through the pattern. This stage is final and it is very hard to remove the formed patterns without causing damage to the underlying wafer.

The overall lithography process is summarized in figure 4.3.3.2 After the pattern is formed on the photoresist and the wafer surface is exposed (*developing process*) the exposed wafer surface is etched. It is also possible to deposit material on the exposed surface.

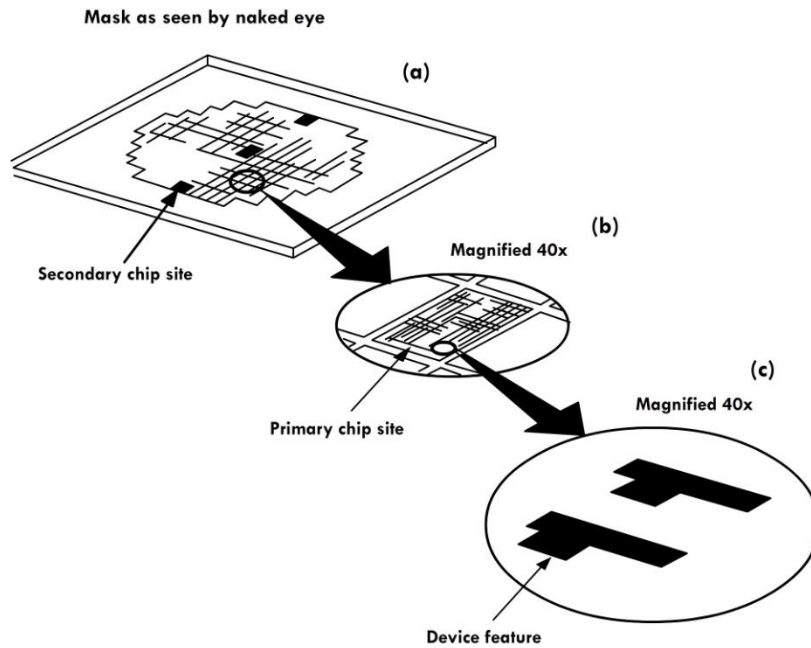


Figure 4.3.3.1.: Typical IC fabrication process showing the different features on the die with increasing magnification from (a) - (c). A mask can be made of many chips, each chip will also have a variety of device features. These patterns will be transferred to the wafer during lithography.

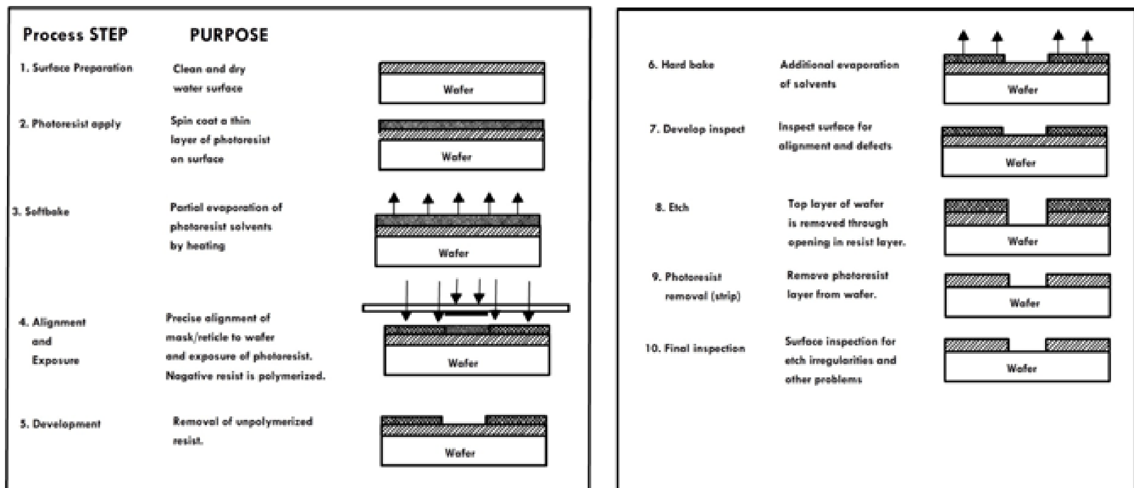


Figure 4.3.3.2: Overview of the lithography process.

4.3.3.3 Photo resists

The use of photo-resists in the wafer fabrication industry was started in the 1950s. The technology was adapted from the photo industry. There are both general purpose resists and resists for specific applications. They are usually tuned to a specific wavelength. The components of a photoresist are as follows.

1. **Polymer** - this is a light sensitive polymer whose structure changes on exposure to light. The desired property is usually change in solubility in a specific solvent.
2. **Solvent** - The solvent is used to thin the resist so that is can be applied on the wafer by a *spin on process*. The solvent is usually removed by heating to around 100 ˚C, called *soft bake process*.
3. **Sensitizers** - these are used to control the chemical reaction during exposure.
4. **Additives** - various chemicals that are added to achieve specific process results, like dyes.

Photoresists usually react to UV or visible light and hence these are called *optical resists*. There are also specific resists for other type of radiations like x-ray and e-beam.Overall, photoresists are divided into two main types.

1. **Positive resists** - on exposure to UV light these become *more soluble*.
2. **Negative resists** - on exposure to UV light these resists becomes *less soluble*.

The difference in working of the two resist types are summarized in figure 4.3.3.3 Positive resists directly transfer the pattern from the mask onto the wafer. This is because the mask protects the portion of the resist below it from exposure to UV radiation. The rest of the resist, that is exposed, becomes more soluble and can be easily removed. Negative resists, on the other hand, transfer the *negative* of the mask pattern to the wafer. This is similar to the negative process in film photography. For negative resists, the portion that is protected by the mask pattern is more soluble, since it is not exposed to UV radiation, while the radiation hardens the rest of the resist.
SU-8 is an example of a commonly used epoxy-based negative photoresist. The structure of the molecule is shown in figure 4.3.3.3.4. It is a viscous polymer based resist. When exposed to UV light of wavelength 365 *nm*, the polymer

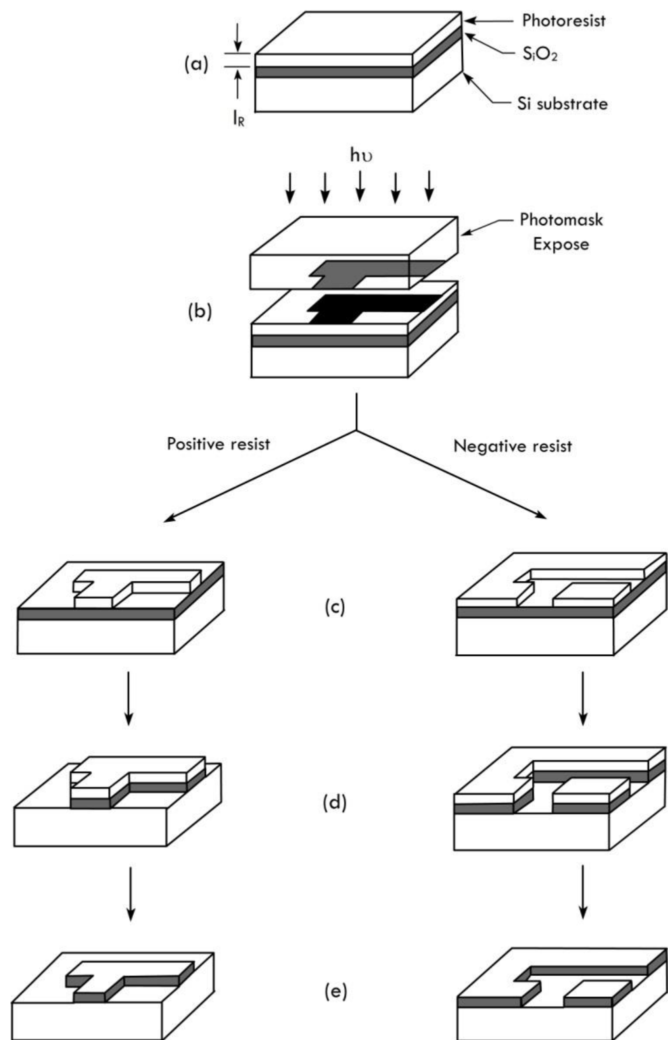


Figure 4.3.3.3: (a) - (e) Steps in exposure of a wafer using positive and negative photoresists with the same mask.

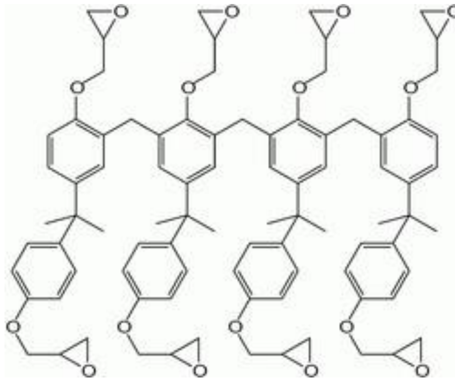
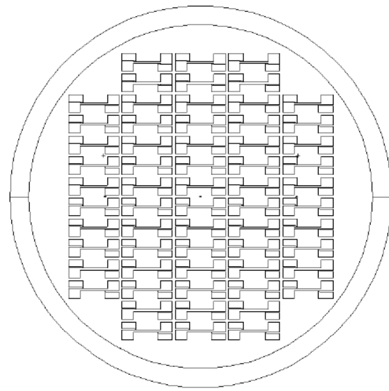


Figure 4.3.3.4: Structure of the SU-8 photoresist.

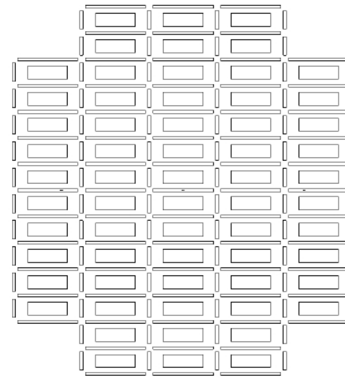
chains cross-link making the resist insoluble. The cured cross-linked chains are stable in vacuum, which is important when using the resist for vapor deposition. Typical photo resist thickness on the wafer is around few hundred *nm* to tens of *μm* depending on the size of the mask pattern. There are a large number of resists and developer groups that are used not only in the IC industry but also for MEMS (micro electro mechanical systems) applications.

4.3.3.4 Mask making

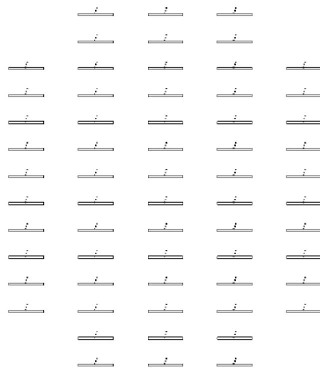
The mask contains the hard copy of the pattern that has to be transferred to the different wafers during lithography. For a given integrated circuit, there are multiple masks, which have to be aligned for proper device fabrication. Masks have *alignment markers* included with the pattern, which can be used for this purpose. Figure 4.3.3.3.5shows three masks used for a MEMS device called nanocalorimeter. The device required three masks, which have to be aligned. This is done by using alignment markers, seen in the center of figure 4.3.3.5 (a) and (b). The alignment markers are usually much smaller than the typical dimensions of the pattern.The mask material is made of borosilicate glass or quartz with a sputter deposited chrome layer on top. The chrome layer is 100 *nm* thick. There is also a photoresist layer deposited on top of the chrome. A laser writer is used to '*write*' the pattern on the mask. Different laser wavelengths (365, 248 or 193 *nm*) and lenses are used to write the pattern on the mask. The choice of the wavelength depends on the smallest dimension on the pattern. The



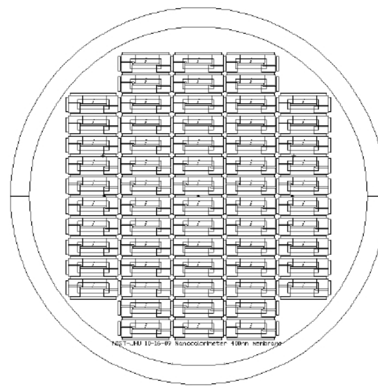
(a) **Layer 1 (front)**



(b) **Layer 2 (back)**



(c) **Layer 3 (top)**



(d) **Combined**

Figure 4.3.3.5: Masks for a nanocalorimeter. (a) Front (b) back (c) top and (d) combined.

laser writing process is sequential (line by line) and can take hours depending on the complexity of the pattern. The mask pattern shown in 4.3.3.5 took approximately 7 hours to write, using a 365 *nm* laser wavelength. After the pattern is written, a suitable developer is used to remove the unexposed photoresist. After that, the exposed chrome layer is removed (using an acid bath etch) and then the remaining photoresist is removed to leave behind the chrome desired pattern on glass. There are also cleaning and drying steps to remove any excess solvent and keep the mask free of dust particles. The major steps in mask making are summarized in figure 4.3.3.6

4.3.3.5 Photoresist application

Before the lithography step, the wafer surface should be clean and defect free. Presence of defects, before and after lithography, can affect the pat- tern transfer process and produce a non-working device. The various ways in which dust particles can interfere with the lithographic mask are shown in figure 4.3.3.7 The dust particles are removed prior to lithography, by washing with de-ionized water, spin drying (rotating the wafer at few thousand *rpm*), hot nitrogen blow-off and a dehydration bake to remove any excess water. The wafers are then inspected for defects and the process repeated, if needed. The photoresist layer is then applied of the wafer. The resist should be uni- formly spread on the surface since any thickness variations can cause prob- lems during developing and subsequent resist removal. Typical resist thick- ness is around 0.5-1.5 μm. Resist application is done by a process called *spin coating*, summarized in figure 4.3.3.8.The photoresist is initially dispensed onto the wafer at rest, called *staticspin coating*. Usually the wafer is held on a vacuum chuck to prevent motion. The chuck is then slowly rotated to spread the photoresist on the surface. This layer is not uniform. After that, the rotation speed is increased to a few thousand *rpm* and the wafer is spun for few tens of seconds, so that excess resist is removed, and there is a uniform film over the entire surface. The right amount of resist should be added, so that coverage is uniform but not excessive, as shown in figure 4.3.3.9 The final resist thickness depends on the amount of resist, spin speed, viscosity, surface tension, and drying char- acteristics (solvent dependent). The relation between resist thickness and spin speed is shown in figure 4.3.3.10 There are other variations to the photore- sist dispersion. The wafer is rotated at slow speed, while resist is dispersed, called *dynamic disperse*. The dispersion arm is moved on the wafer surface

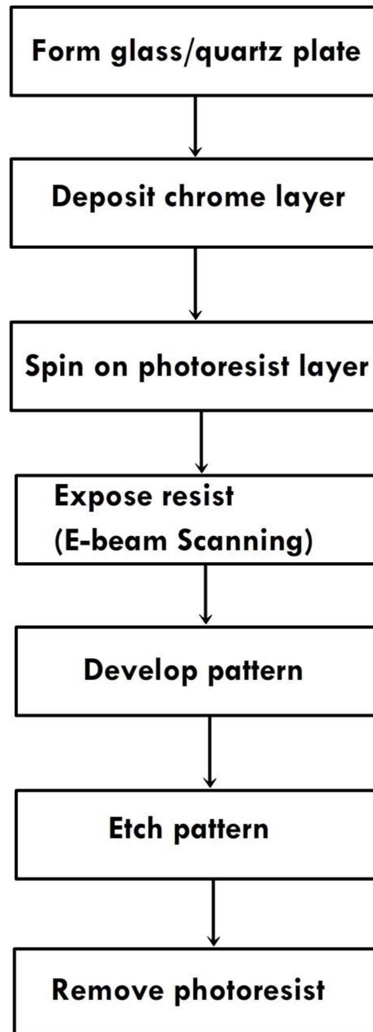


Figure 4.3.3.6: Process flow for the mask making process.

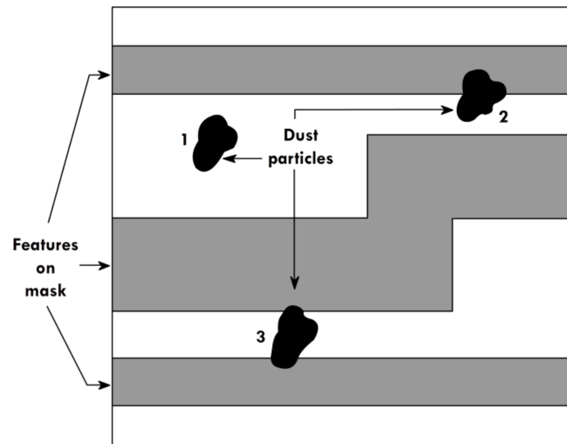


Figure 4.3.3.7: Dust particles can interfere with the lithography process and cause errors in the pattern transfer.

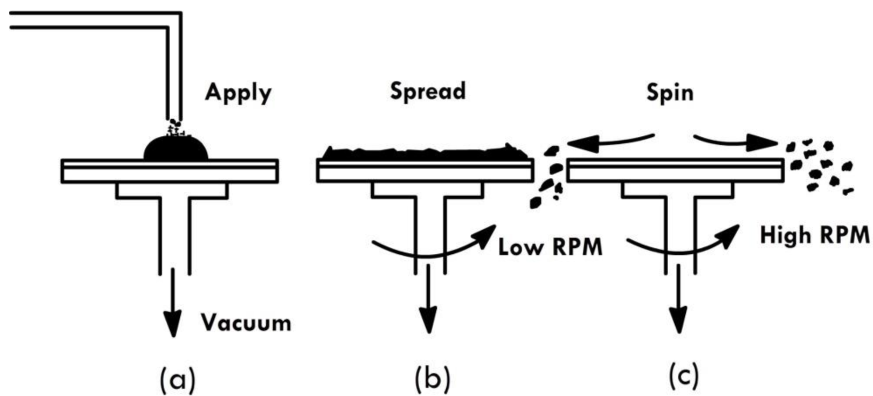


Figure 4.3.3.8: Steps in spin coating to get a uniform layer of resist. (a) A layer of resist is first applied on the wafer (b) The wafer is rotated at low rpm to spread the resist (c) The wafer is spun at high rpm so that an uniform coating is obtained and excess resist removed.

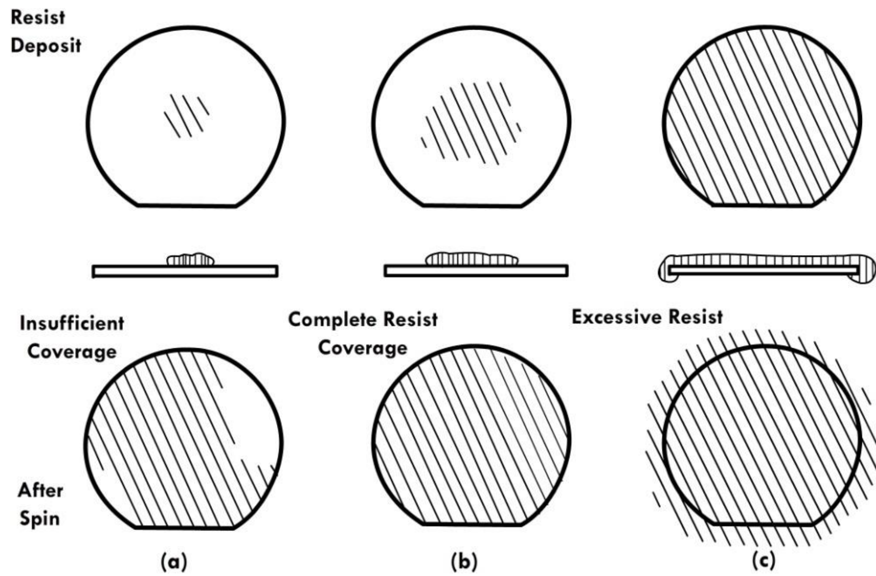


Figure 4.3.3.9: Resist coverage before and after 'spinning' for (a) insufficient resist (b) Correct amount of resist and (c) excess resist. Resist dispensing is usually an automated process.

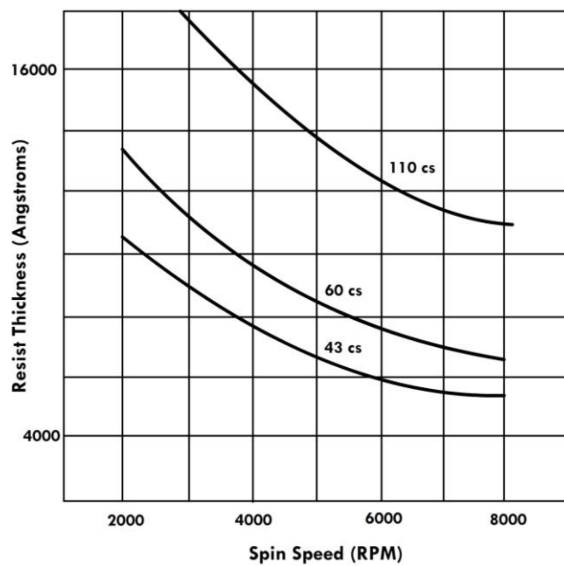


Figure 4.3.3.10: Resist thickness vs. spin speed for different volumes of resist dispersed on the wafers. The resist thickness increases with the volume of material dispensed. Also, as spin speed increases the thickness decreases.

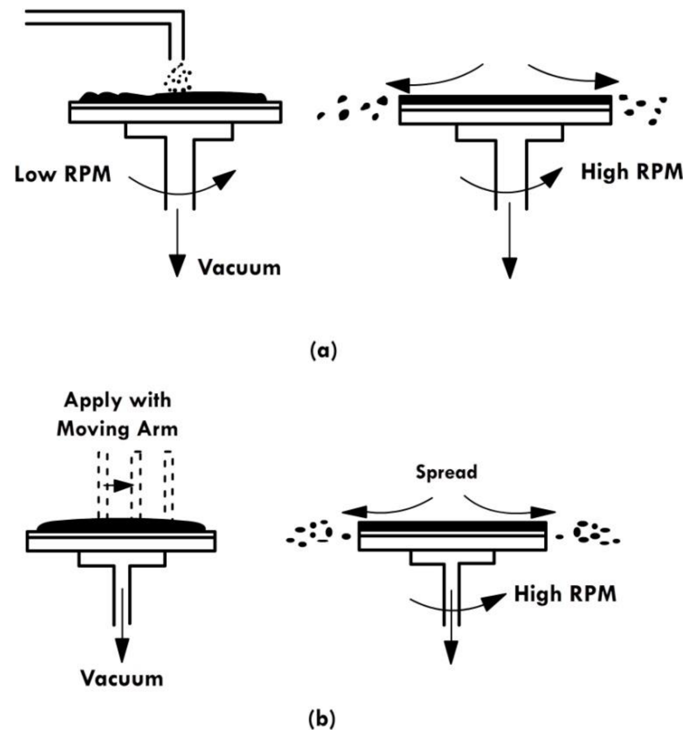


Figure 4.3.3.11: (a) Dynamic disperse (b) Moving arm disperse. Both are used to achieve uniform coverage, especially for large wafers used in commercial IC fabrication.

while dispersion, called *moving arm disperse*. All these different techniques are used to achieve uniform coverage, especially for large wafers. The above mentioned techniques are summarized in figure 4.3.3.11.The photoresist application process is automated in commercial IC man- ufacturing. In most research based facilities, for small (3”-4”) wafers, the dispersal is usually manual. After spinning, the wafer is subjected to a *soft bake* process. This heats the wafer to 100-120 ˚C to remove the solvent from the resist. After spin on process, the wafer surface should be protected from ambient light (typically UV light) to prevent *unintentional exposure* of the resist. This is done by keeping the photoresist application under special light- ing conditions. The alignment and exposure system is usually kept close to the spin on process equipment to minimize exposure.

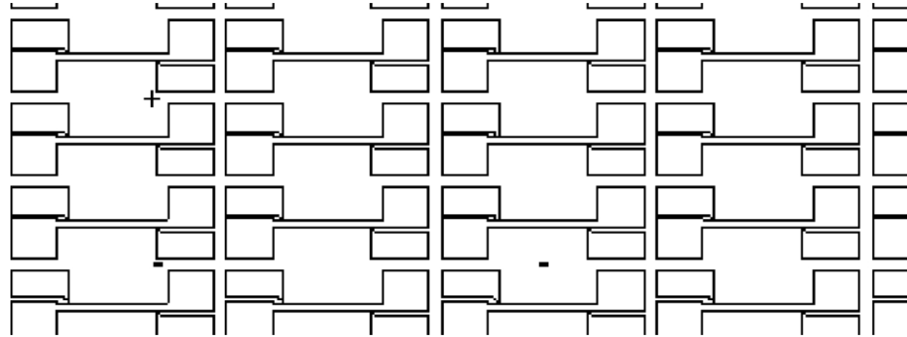


Figure 4.3.3.12: Alignment marks for the mask shown in figure 4.3.3.5 (a). The marker is located at the center of the mask region.

4.3.3.6 Alignment and exposure

The alignment and exposure process transfers the pattern from the mask to the photoresist on the wafers. *Alignment markers* are used to align the mask with the wafer and also to align one more masks with each other. Figure 4.3.3.12 shows alignment markers for the mask shown in figure 4.3.3.5 (a). The pattern is transferred from the mask to the photoresist using *steppers*. The transfer can be 1:1 i.e. direct transfer of the pattern onto the wafer. There are also *reduction steppers*, where the reticles can be 5-10 times larger than the final dimensions on the wafer. In such cases, the reticle is projected onto one area of the wafer and then *stepped* to the next area. The advantage is that smaller dimensions can be achieved by using a larger mask.
The stepper can be of a contact type, where the mask actually touches the wafer or a proximity type, where there is a gap. These types are shown in figure 4.3.3.3.13 Contact aligners can cause damage to the mask (since they have to repeatedly used on different wafers) and have contamination issues. So proximity aligners are preferable, though there is a slight loss of resolution due to scattering of light in the gap. Some sort of soft contact contact aligners are also available. There are different modes of projection, as shown in figure 4.3.3.14

4.3.3.7 E-beam lithography

In conventional lithography, a laser writer is used to create a hard copy of the pattern i.e. mask, which is then transferred to the wafers. The size limitation comes from the smallest features that can be written and this depends on the wavelength of light used (few hundred *nm*). One way to circumvent this limitation

is to use an electron beam, since this has a much

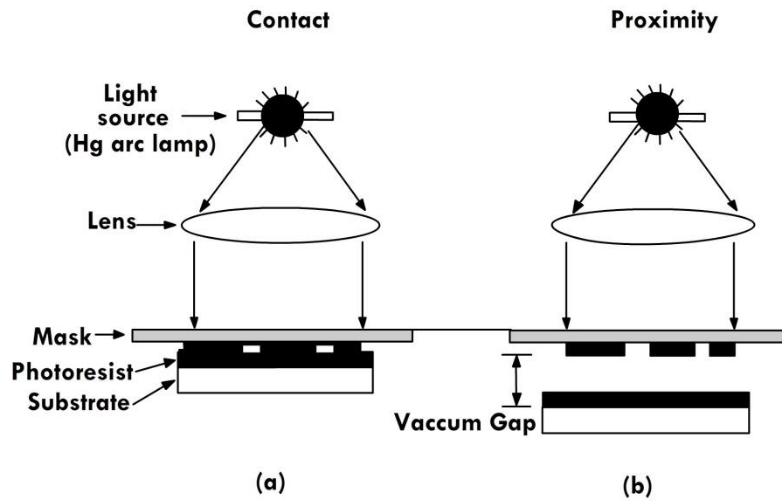


Figure 4.3.3.13: Types of stepper (a) contact (b) proximity.

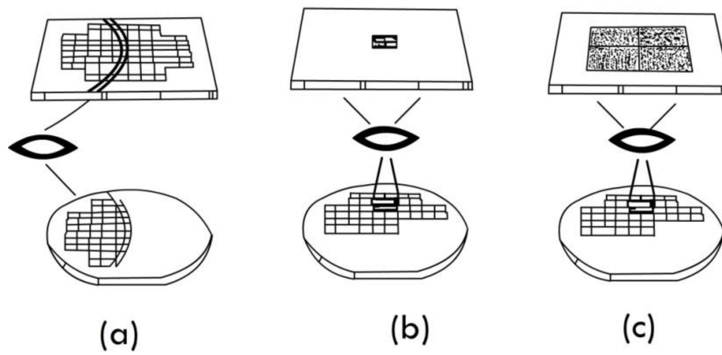


Figure 4.3.3.13: Types of projection systems (a) Scan (b) 1:1 step and repeat (c) reduction step and repeat.

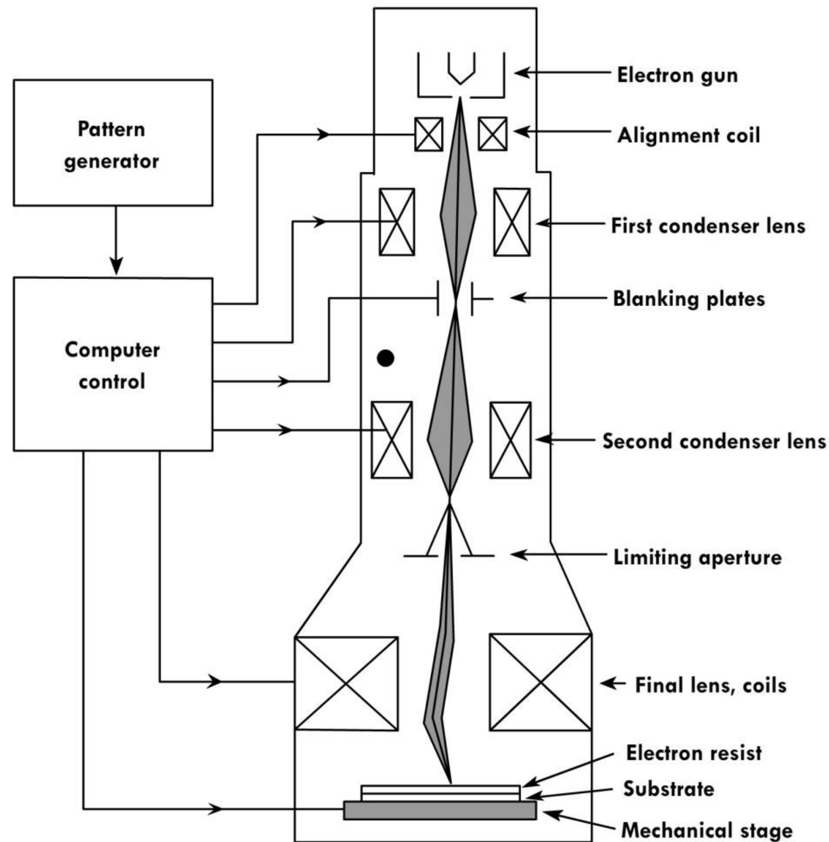


Figure 4.3.3.14: Electron beam lithography setup.

smaller wavelength (few *nm* depending on energy) and hence can *theoretically* achieve a much higher resolution. In e-beam lithography, the electron beam is used to scan and write the design directly on the wafer. This is called *direct writing*. The setup is shown in figure 4.3.3.15. It is similar to a scanning electron microscopy setup, with an electron source and lens and deflector coils to scan the beam on the surface. Resolution better than conventional optical lithography can be achieved, but the disadvantage is that each wafer has to be *written individually* and the process is time-consuming. Also, e- beam lithography is a scanning system while conventional lithography is a one shot exposure system.

4.3.3.8 Developing

After the alignment and exposure process, the wafers have to be *developed*. The terminology is similar to that used in film photography. The wafers are reacted with a suitable chemical (developer) that reacts with the exposed photoresist. The type of developer chosen depends on the resist. For a positive photoresist, the exposed areas are removed (more soluble) while for a negative resist, the unexposed areas are removed (less soluble). SU-8 is a negative photoresist, whose structure shown in figure 4.3.3.4. After exposure, the main developer used to remove the unexposed resist is *1-methoxy-2-propanol acetate*.Developing is usually a wet chemical process. The wafers are immersed in the developing solution for a fixed time, until the resist is completely removed. They are then cleaned and dried. After that, the wafers are baked to 200- 250 ˚C, called *hard bake*, to harden the remaining resist. At this stage, the pattern that needs to be transferred to the wafer is still only temporary. It is possible to remove the resist easily, usually by dry etching. The developed wafers are then further processed to get the final pattern on the wafer. These could include steps like

a. Doping - ion implantation only. For thermal diffusion, oxide layers are used as masks.

b. Deposition - usually a physical vapor deposition process like sputtering or e-beam evaporation. Chemical vapor deposition can react with the wafers.

c. Etching - plasma or reactive ion etching. Wet etching can damage the remaining resists.

The resist protects the portion of the wafer that lies below it. After the final pattern is obtained on the wafer, the remaining resist is removed, this is called **resist stripping**. This can be a *wet process*, by using an acid mixture or a *dry process*, plasma etching with oxygen. The wafers are then cleaned and dried and are ready for the next process. If there are multiple lithography steps, the wafers then go back to the photoresist application process.

4.3.4 Etching

4.3.4.1 Introduction

Etching refers to the removal of material from the wafer surface. The process is usually combined with lithography in order to select specific areas on the wafer from which material is to be removed. Etching represents one way of permanently transferring the mask pattern from the photoresist to the wafer surface. The complementary process to etching is *deposition* (or growth), where new material

is added. Unlike oxidation (or nitridation), where the underlying Si is consumed to form the oxide (nitride) layer, in deposition, new material is added without consuming the underlying wafer.

4.3.4.2 Etching challenges

There are some process challenges related to etching. These are common to both wet and dry etching, though they are more pronounced and harder to control in wet etching due to the higher rate of material removal, compared to dry etching.In incomplete etch, the time is not sufficient for complete material removal. This is usually due to concentration or temperature not being sufficient. The concentration profile left behind is usually a rough surface, due to local variations in material removal, as depicted in figure 4.3.4.5.

4.3.4.3 Wet etching

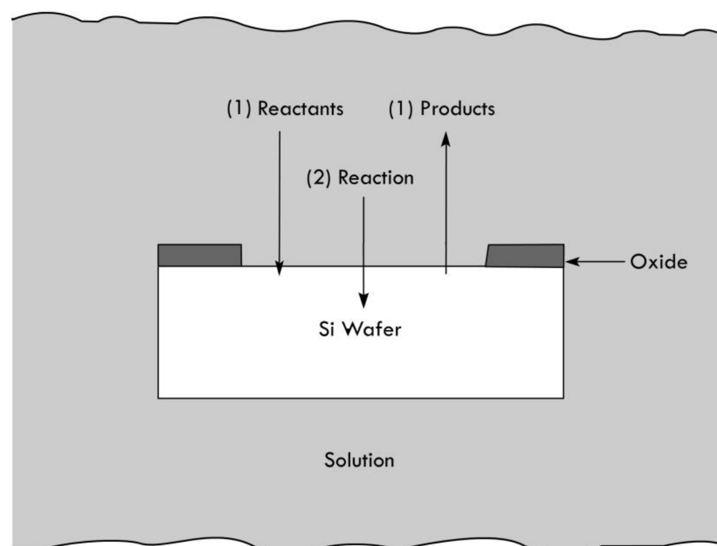


Figure 4.3.4.1: Schematic of the wet etching process. A controlled portion of the wafer surface is exposed to the etchant which then removes materials by chemical reaction

In wet etching, the wafers are immersed in a tank of the etchant (mix of chemicals), as shown in figure 4.3.4.1. There is a chemical reaction between the wafer surface and the etchants that helps in material removal. Either a photoresist layer or a hard mask like oxide or nitride layer is used to protect the rest of the wafer. The time for etching depends on the amount and type of material that needs to be removed. KOH (potassium hydroxide) is a common etchant used to remove Si. Usually, 30% KOH solution is used, which has a etch rate of 100 *μm/hr* at 90 ˚C. Thus, an entire 4" wafer, with thickness of 500 *μm*, can be etched through in approximately 5 hours. The etch rate of Si (100) by 30 % KOH is shown in figure 4.3.4.2. After etching, the wafers are rinsed, usually in DI water, for removal of etchant and then finally dried.Wet etching is used for removal of material from large areas (trench sizes > 3 *μm*). For smaller areas, where greater precision in removal of material is required, dry etch is preferred. The wet etching process is anisotropic i.e. the etch rate depends on the plane of the Si wafer, from which atoms are being

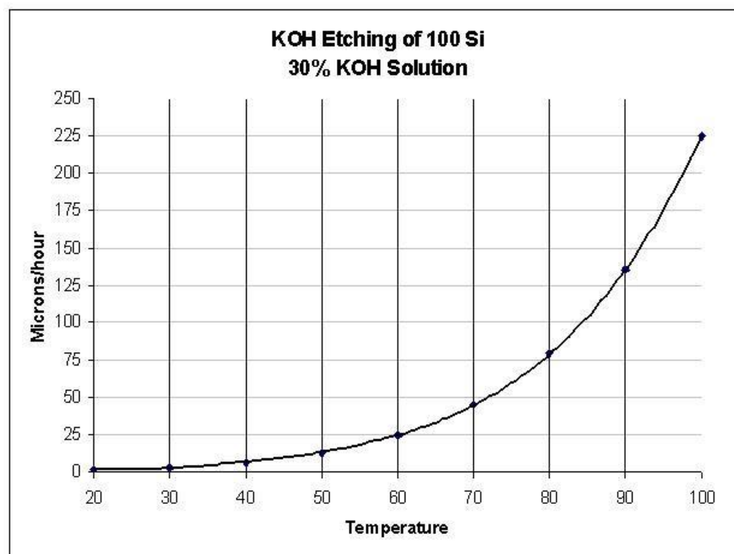


Figure 4.3.4.2: Etch rate of Si in KOH as a function of temperature. There is a non-linear increase in etch rate with increasing temperature.

removed. The etch rate for Si (110), in the same 30 % KOH, is shown in figure 4.3.4.3. Compared to Si(100) plane, figure 4.3.4.2, the rate is higher. This means that wet etching of Si(100) will produce a trapezoidal profile, with a specific angle of 54.74 deg,

as shown in figure 4.3.4.4. Etching uniformity is important to get a uniform thickness over the entire wafer surface. This is usually determined by process conditions like etchant temperature, concentration, and agitation (using stirrers).

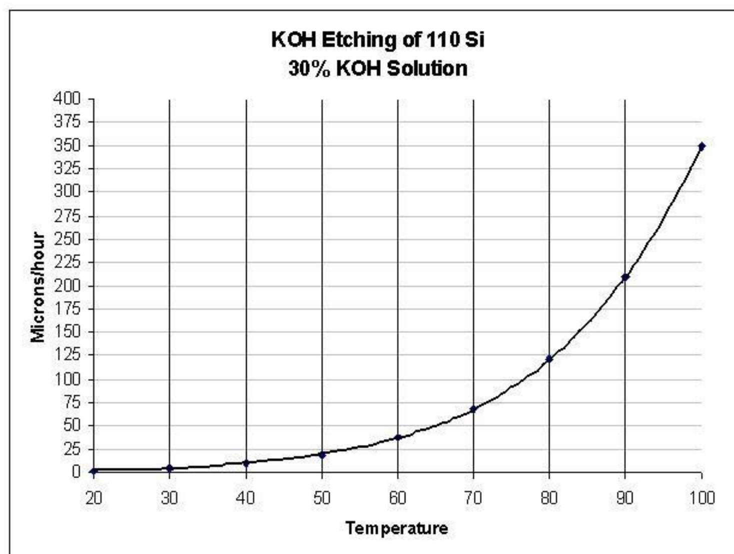


Figure 4.3.4.3: Etch rate of Si(110) in KOH as a function of temperature. Compared to the etch rate for Si (100),

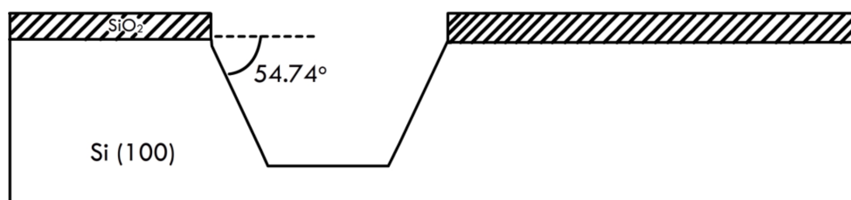


Figure 4.3.4.4: Anisotropic etching of Si by KOH.

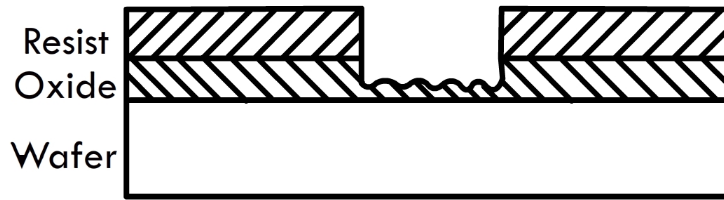


Figure 4.3.4.5: Surface profile for incomplete etch of oxide layer on Si. A resist layer protects the remaining oxide. A rough oxide layer is left behind due to the local variations in rate of removal of the oxide layer..

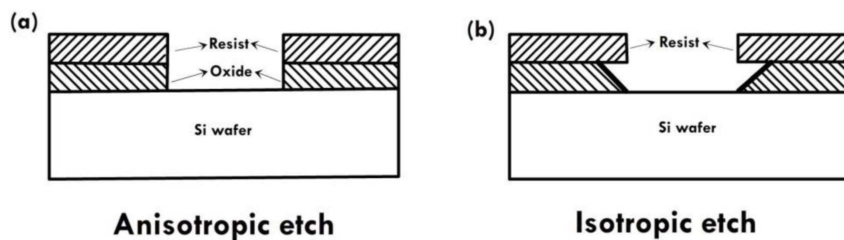


Figure 4.3.4.6: (a) A complete anisotropic etching produces vertical side walls. (b) Most often etching is partially isotropic, so that side walls are formed at an angle.

The opposite of incomplete etching is over etching. An ideal etchant is se- lective and completely anisotropic. This is essential to get vertical sidewalls when a trench is created. But, this is not always possible, so that sloped side walls are obtained, see figure 4.3.4.6. When the etch time is larger than the required etch time, due to isotropic etching, material under the photoresist can get removed. This is called *over etching* and in extreme cases it can also lead to liftoff of the resist layer, see figure 4.3.4.7. This is harmful, since it exposes areas of the wafer that the resist protects to the etching process.Etching process should be selective to the material that has to be removed. This helps to protect the material under the mask (within limits of isotropic etching) and also the mask material itself (oxide, nitride, or resists). Consider

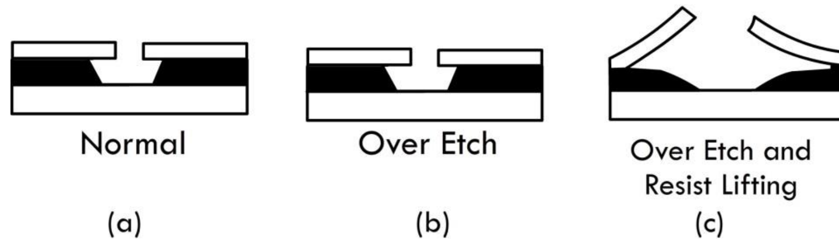


Figure 4.3.4.7: (a) Normal (b) Over etching and (c) Resist liftoff due to excess over etching. The etching rate and time are crucial to prevent over etching since resist removal can cause damage to portions of the wafer that have to be protected from the etchant.

Si etching, using KOH. The etch rate for Si(100) at 90 ˚C using 30 % KOH is 100 *µm/hr*, see figure 4.3.4.2. If silicon nitride is used as mask, its etch rate, under the same conditions, is 1 *nm/hr*, nearly 10^5 times slower than Si. Thus, silicon nitride is commonly used as a mask for Si etching (especially for making Si cantilever based devices). On the other hand, silicon dioxide etch rate, under the same conditions, is 1 *µm/hr*, see figure 4.3.4.8. So, using silicon oxide as a mask will not be good enough or a very thick oxide layer is required.Different etchants that are used for different layers and the corresponding etch rates, shown in table1. For Si etching, KOH is used or a mixture of nitric acid and hydrofluoric acid (HF). For silicon oxide etching, usually a mixture of HF and ammonium fluoride (NH_4F) is used, that produces a etch rate of 0.1 *µm/hr* at room temperature. This mixture does not etch Si, so it provides very good selectivity. This etchant is called *BOE* (buffered oxide etchants). For silicon nitride, usually a strong acid like hot phosphoric acid is used at high temperatures (180 ˚C) since it is a very good passivating layer and hard to remove under normal conditions.

4.3.4.4 Dry etching

Dry etching, as the name suggest, is removal of material in the absence of solvent. The process was introduced because wet etching has some limitations in its applicability, which are listed below.

1.Wet etching is used for large pattern sizes, usually larger than 2 µm. 2.It is an

isotropic process - sloped sidewalls rather than straight walls.

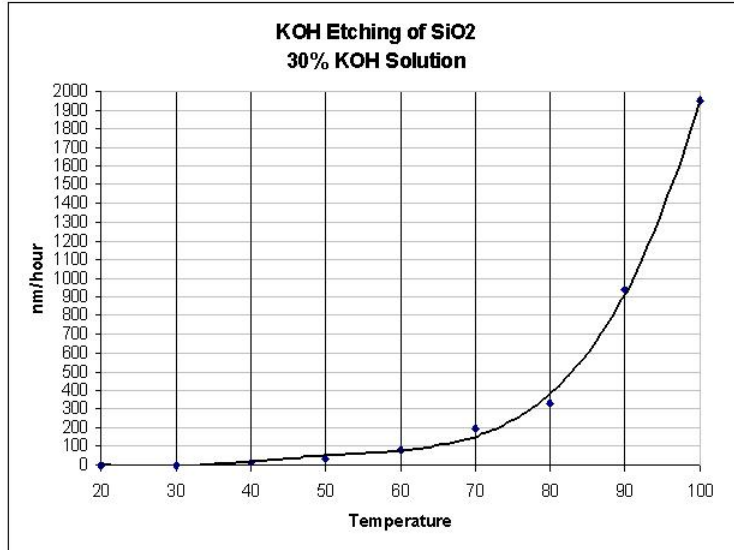


Figure 4.3.4.8: Etch rate of SiO_2 by KOH.

(Because of comparable etch rates to pure Si, it cannot be used as a etch mask for Si etching. Silicon nitride is used as the mask since its etch rate, under the same conditions, is of the order of *nm/hr*)

Table 1: Etching chemicals used for different layers and their etch rates, under commonly used conditions. Adapted from *Microchip fabrication - Peter van Zant.*

Material	**Common etchant**	**Etch temperature**	**Etch rate** (*Å/min*)
SiO_2	HF NH_4F (1:8)	Room temperature	700
SiO_2	Acetic acid NH_4F (2:1)	Room temperature	1000
Aluminum	HPO_4 HNO (nitroxyl) Acetic acid water	40-50 ˚C	2000

Si_3N_4	H_3PO_4	150-180 °C	80
Poly Si	HNO_3 H_2O HF	Room temperature	1000

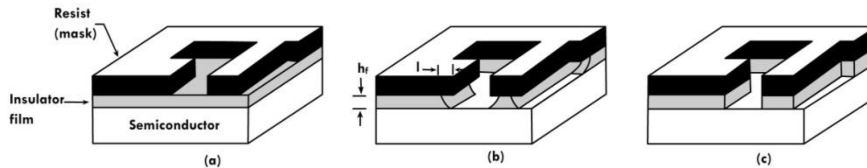


Figure 4.3.4.9: (a) Starting surface after development of the resist (b) Surface after wet etching (c) Surface after dry etching. Because of the anisotropic nature of etching, dry etching produces more vertical side walls compared to wet etching, but the removal rate is slower. Adapted from *Fundamentals of semiconductor manufacturing and process control - May and Spanos.*

3. Wet etch has to be combined with subsequent rinse and dry steps. This increases chances of defects or contamination.
4. Hazardous chemicals and conditions are used, so safety is an issue. Safe disposal of chemicals is essential.
5. Undercutting and resist peel off can happen if time is not controlled or etch conditions change during process.

The wet and dry etching process are compared in figure 4.3.4.9. Dry etching is a process that overcomes some of these issues. Here, *etchant gases* are the primary medium for the removal of material. The basic steps involved are summarized in figure 4.3.4.10. There are three main types of dry etching.They are Plasma etch,Ion beam etching and reactive ion beam etching

4.3.4.4.1 Plasma etch

In plasma etch, the chemical etchant is introduced in the gas phase. For etching silicon oxide, CF_4 (tetrafluoromethane) is used. The chamber is first evacuated before introducing the gas. Radio frequency (RF) electrodes are then used to generate the plasma that ionizes the gas. This ionized gas attacks the oxide layer,

removing the layer. Etch rates in plasma etch are
~ 1 ±10 $\mu m/hr$, much smaller than wet etching. So, it more suitable for thin layers, but it also provides greater thickness control. There are different configurations for plasma etching, one such *planar configuration* is shown in

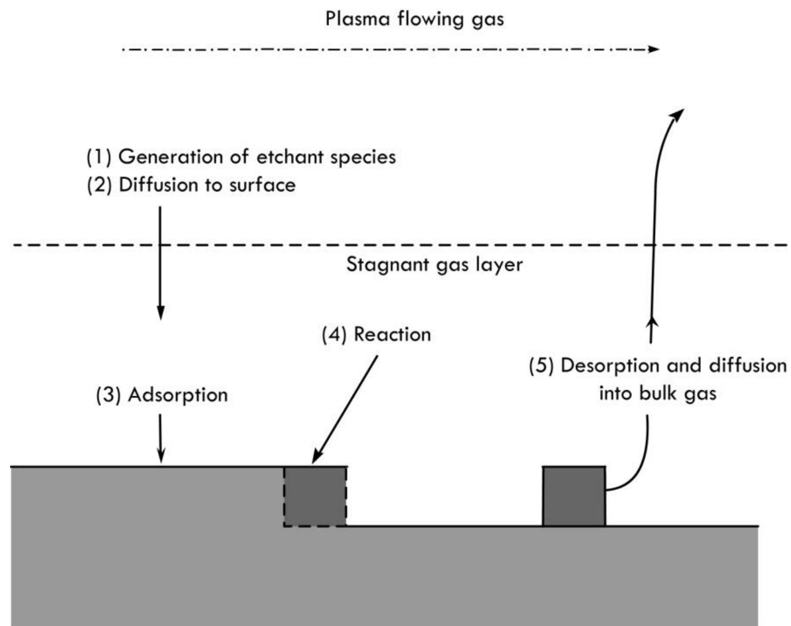


Figure 4.3.4.10: Various steps from (1) - (5) in the dry etch process

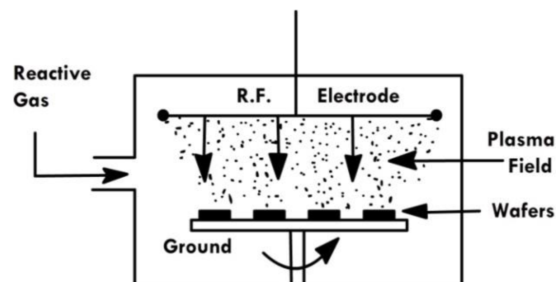


Figure 4.3.4.11: Planar plasma etch configuration.

Table 2: Typical plasma etching chemicals for different film materials and the corresponding gaseous products. Adapted from *Microchip fabrication - Peter van Zant.*

Film	Etchant	Typical gas compounds
Al	Chlorine	BCl_3, CCl_4, Cl_2, $SiCl_4$
Mo	Fluorine	CF_4, SF_4, SF_8
Polymers	Oxygen	DF_4, SF_4, SF_8
Si	Chlorine Fluorine	BCl_3, CCl_4, Cl_2, $SiCl_4$ CF_4, SF_4, SF_6
SiO_2	Fluorine	CF_4, CHF_3, C_2F_6, C_3F_8
Ta	Fluorine	CF_4, CHF_3, C_2F_6, C_3F_8
Ti	Fluorine	CF_4, CHF_3, C_2F_6, C_3F_8
W	Fluorine	CF_4, CHF_3, C_2F_6, C_3F_8

figure 4.3.4.11. The resist layer used to protect the wafer is also etched along with the oxide. But the resist thickness is much larger than the oxide (few μm of resist compared to tens of *nm* of oxide). This means that substantial amount of resist is still available, after the etching process. Some of the different etchant gases used for plasma etching of various films are shown in table 2.

4.3.4.4.2 Ion beam etch

Ion beam etching is similar to the ion beam milling process that is used for transmission electron microscopy sample preparation. This is a physical process where ionized inert gas ions (usually Ar) are used to remove material from the wafer. The process is *not selective* but it is *highly directional*. The ion beam etching process is shown in figure 4.3.4.12.

4.3.4.4.3 Reactive ion etching

Reactive ion etching combines the plasma and ion beam etching process to achieve both selectivity and directionality. There is an increase in selectivity compared to plasma etch, for SiO_2 and Si the selectivity ratio is 35:1 while for pure plasma etch the ratio is 10:1. This reduces the thickness requirement on the mask. Dry etch process is also used for resist stripping after patterning is complete. This is usually done by plasma etching using oxygen.

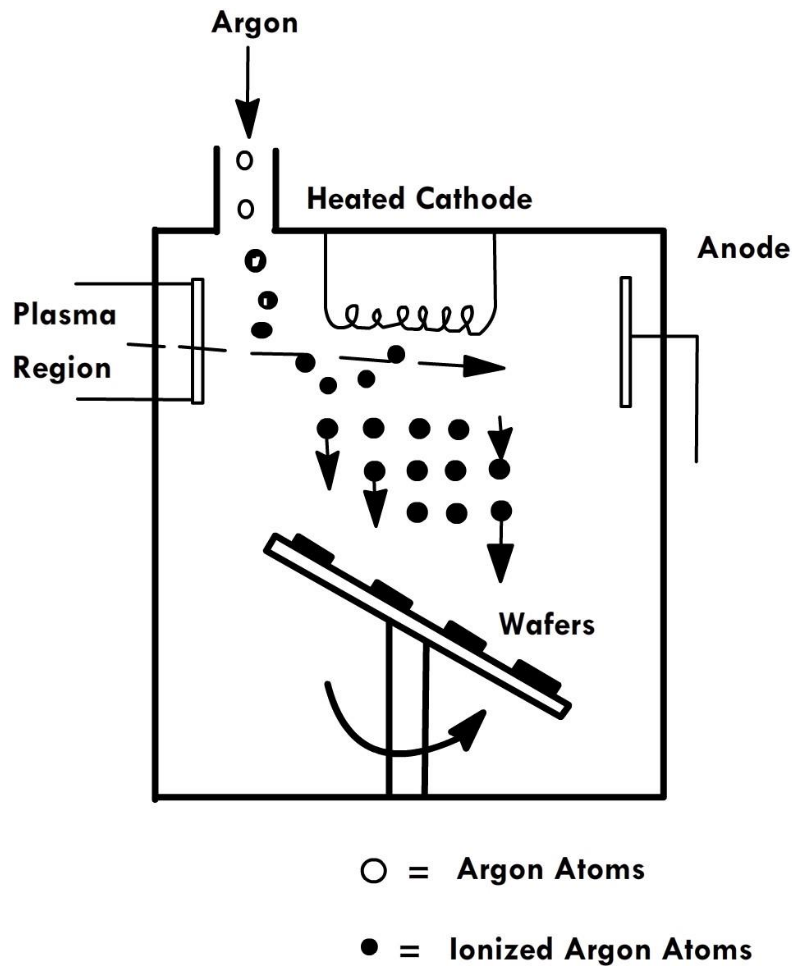


Figure 4.3.4.12: Schematic of the ion beam etching process

4.3.5 Diffusion

4.3.5.1 Introduction

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.

4.3.5.2 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

4.3.5.3 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.

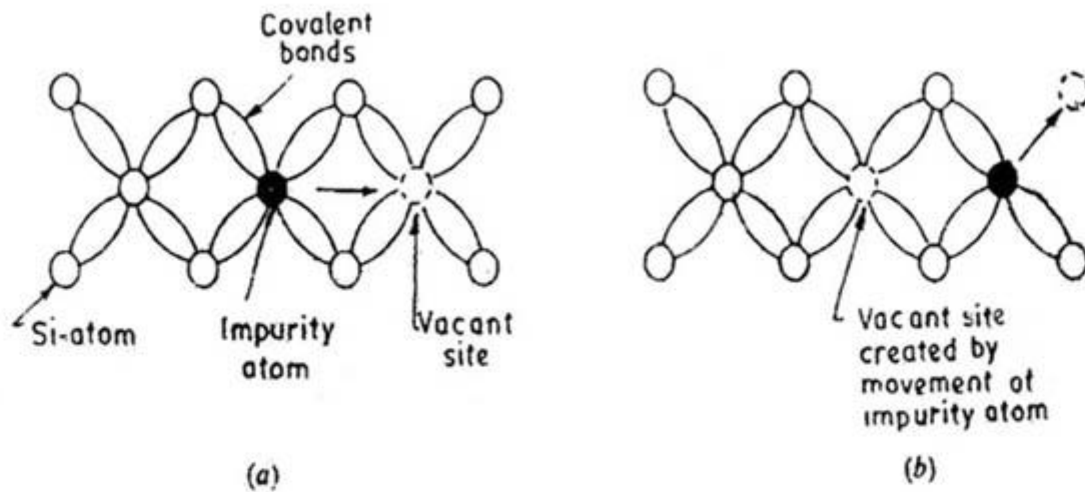


Figure 4.3.5.1 Substitutional Diffusion By Dopant Impurities

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.

4.3.5.4 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.

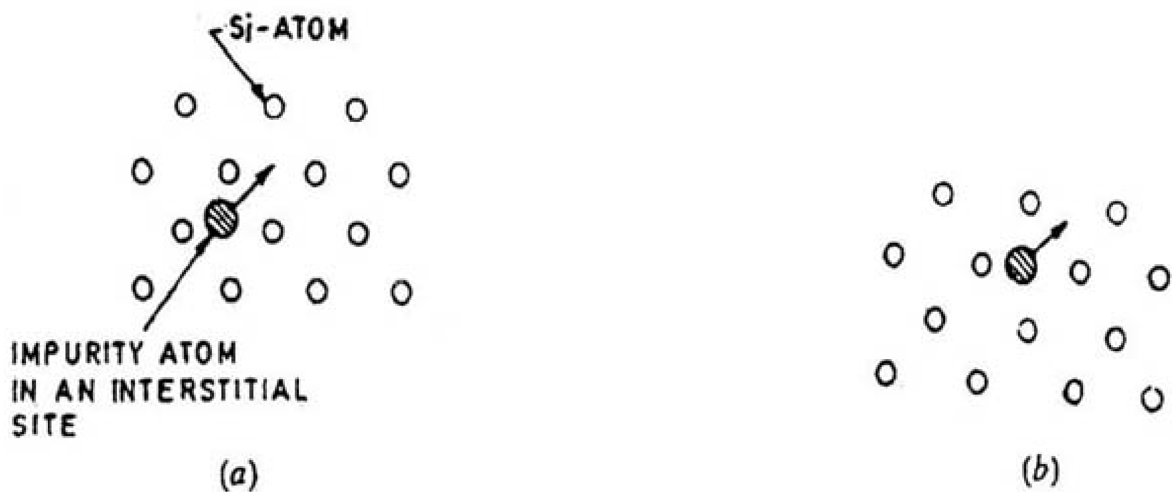


Figure 4.3.5 2 Interstitial Diffusion of Impurity Atom

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.

4.3.5.5 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

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

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,

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

This is Ficks second law of distribution.

4.3.5.6 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.

4.3.5..7 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,

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

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

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

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.

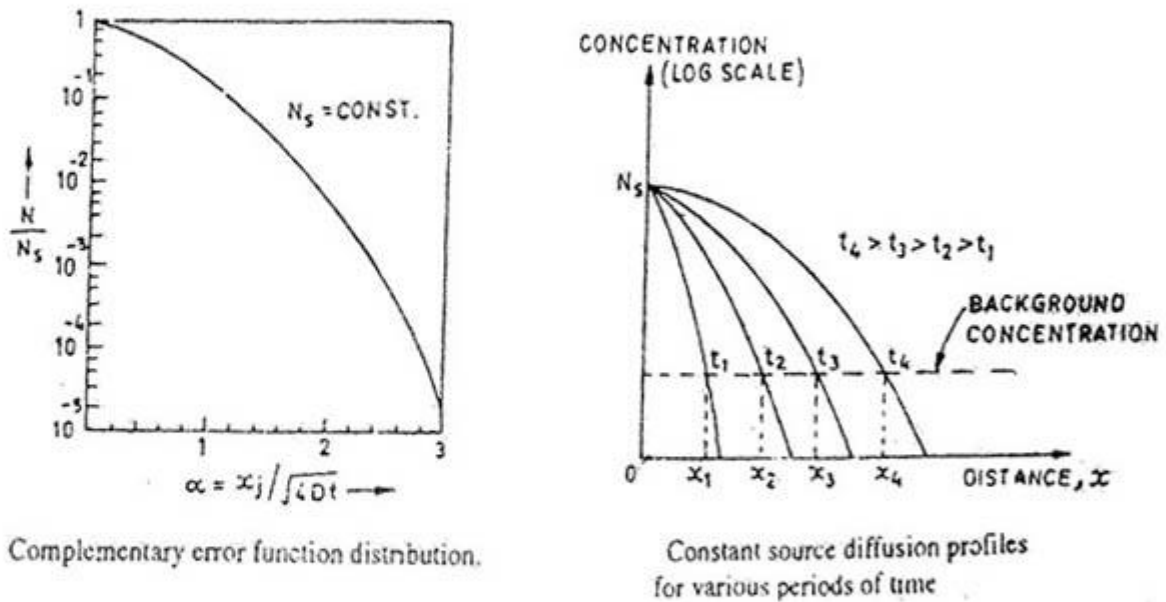


Figure 4.3.5 3 Complimentary Error Function

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.

4.3.5.8 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.

4.3.5.9 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.

4.3.5.10 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.

4.3.5.11 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.

4.3.5.11.1 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.

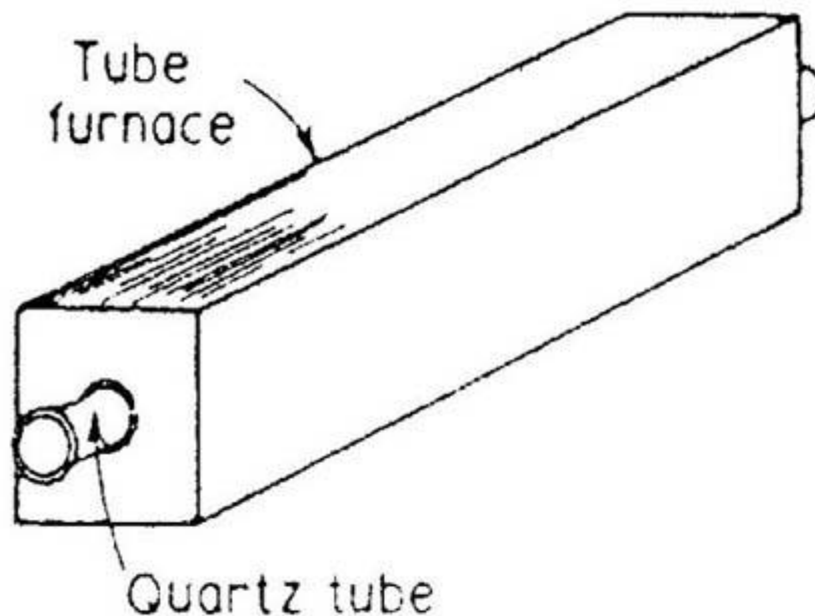


Figure 4.3.5.4 Diffusion Furnace

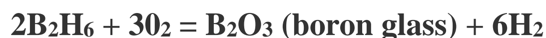
The temperature of the furnace is kept about1000°C. The temperature within the quartz furnace tube can be controlled very accurately such that a temperature within 1/2°C of the set-point temperature can be maintained uniformly over a “hot zone” about 1 m in length. This is achieved by three individually 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:



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.

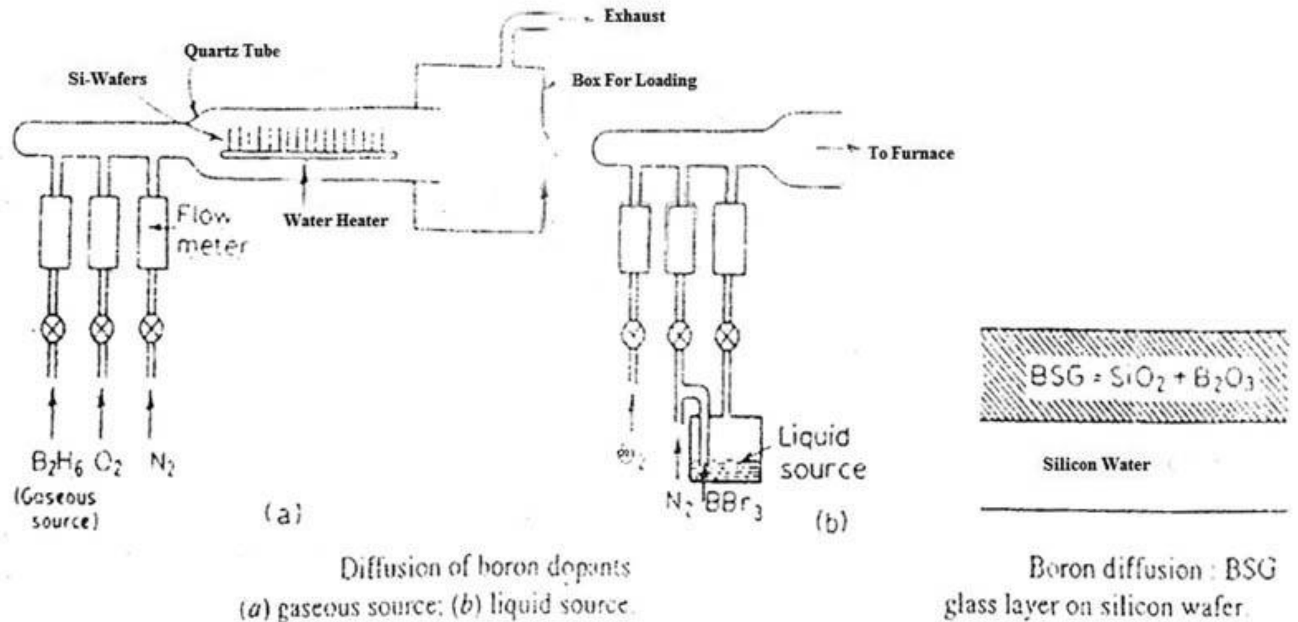


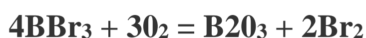
Figure 4.3.5.5 Diffusion Of Dopants

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 :

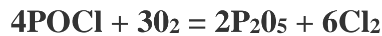


$4BBr_3 + 30_2 = B20_3 + 2Br_2$

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 concentrations and better control of impurity density from wafer to wafer and from run to run.

4.3.6 Ion Implantation

4.3.6.1.Introduction

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 System

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

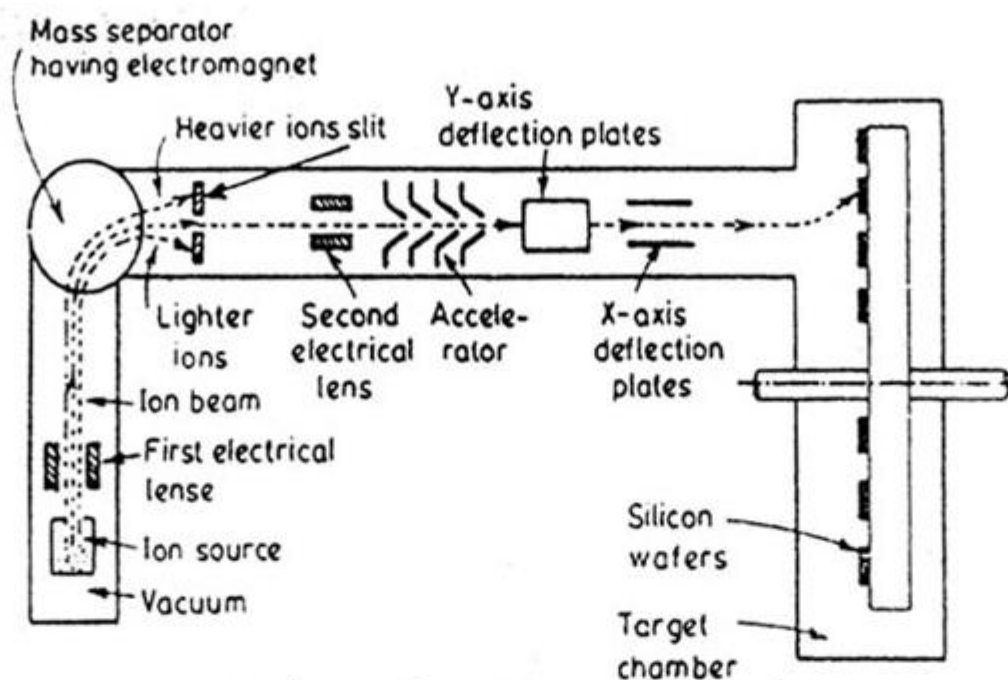


Figure 4.3.6.1 Ion Implantation System

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. For this purpose the focussed

ion beam is scanned electrostatically over the surface of the wafer in the target chamber.Repetitive scanning in a raster pattern provides exceptionally uniform doping of the wafer surface. The target chamber commonly includes automatic wafer handling facilities to speed up the process of implanting many wafers per hour.

4.3.6.2 Properties of Ion Implantation

The depth of penetration of any particular type of ion will increase with increasing accelerating voltage. The penetration depth will generally be in the range of 0.1 to 1.0 micro meters.

4.3.6.3 Annealing after Implantation

After the ions have been implanted they are lodged principally in interstitial positions in the silicon crystal structure, and the surface region into which the implantation has taken place will be heavily damaged by the impact of the high-energy ions. The disarray of silicon atoms in the surface region is often to the extent that this region is no longer crystalline in structure, but rather amorphous. To restore this surface region back to a well-ordered crystalline state and to allow the implanted ions to go into substitutional sites in the crystal structure, the wafer must be subjected to an annealing process. The annealing process usually involves the heating of the wafers to some elevated temperature often in the range of 1000°C for a suitable length of time such as 30 minutes.

Laser beam and electron-beam annealing are also employed. In such annealing techniques only the surface region of the wafer is heated and re-crystallized. An ion implantation process is often followed by a conventional-type drive-in diffusion, in which case the annealing process will occur as part of the drive-in diffusion.

Ion implantation is a substantially more expensive process than conventional deposition diffusion, both in terms of the cost of the equipment and the throughput, it does, however, offer following advantages.

4.3.6.4 Advantages of Ion Implantation

Ion implantation provides much more precise control over the density of dopants deposited into the wafer, and hence the sheet resistance. This is possible because both the accelerating voltage and the ion beam current are electrically controlled outside of the apparatus in which the implants occur. Also since the beam current can be measured accurately during implantation, a precise quantity of impurity can be introduced. Tins control over doping level, along with the uniformity of the implant over the wafer surface, make ion implantation attractive for the IC fabrication, since this causes significant improvement in the quality of an IC.

Due to precise control over doping concentration, it is possible to have very low values of dosage so that very large values of sheet resistance can be obtained. These high sheet resistance values are useful for obtaining large-value resistors for ICs. Very low-dosage, low-energy implantations are also used for the adjustment of the threshold voltage of MOSFET's and other applications.

An obvious advantage of implantation is that it can be done at relatively low temperatures, this means that doped layers can be implanted without disturbing previously diffused regions. This means a lesser tendency for lateral spreading.

4.3.6.5 High-Current High-Energy Implantation Machines

The ion-implantation apparatus, discussed above, has limits to energy range. The minimum implantation energy is usually set by the extraction voltage, that is, the voltage causing the ions to move out of the ion source into the mass separator. This voltage (which is typically 20 KeV) cannot be reduced too far without drastically reducing beam current. The maximum implantation energy is set by the design of the high voltage equipment. The only way to circumvent this is to implant multiply-charged ions.

High beam currents are obtained by using multiple extraction electrodes and higher voltages. To get a final beam of suitable energy a combination of acceleration and deceleration modes of operation is used.

The electrostatic scanning is not suitable for high-beam currents, as it disrupts space charge neutrality and leads to beam "blow-up". Therefore a mechanical scanning system is usually used. In this case, the wafer is scanned past a stationary beam. This method has the added advantage of keeping the same beam angle across the whole wafer, whereas an electrostatic system can vary by ±2° for 100 mm wafers. However, mechanical scanning puts new requirements on the wafer holder.

High-energy implantation, at MeV energies, makes possible several new processing techniques required for VLSI.

High-energy implantation machines however introduce high-voltage breakdown problem. At about 400 KeV of energy electrical breakdown of the air around the high voltage equipment occurs. Hence, above 400 KeV, conventional equipment is used. Also, high energy implants frequently require water stages heated up to 600 degree Celsius, so that self annealing during implantation minimizes damage in the surface layer. Mechanical scanning is used because of the difficulty of electrostatically scanning a high-energy beam.

4.3.6.6 Problems in VLSI Processing

Now a day's large diameter wafers are feasible. Large size wafers are necessary for VLSI. This makes the task of uniformly implanting a wafer increasingly difficult. This in turn has effect on sheet resistance. Ion implantation is basically clean process because contaminant ions are separated from the beam before they hit the target. There are still several sources of contamination possible near the end of the beam line, which can result in contaminant dose up to 10 percent of the intended ion dose, for example, metal atoms knocked from chamber walls, water holder, masking aperature and so on.

Annealing, as discussed earlier, is required to repair lattice damage and put dopant atoms on substitutional site where they will be electrically active. The success of annealing is often measured in terms of the fraction of the dopant that is electrically active, as found experimentally using a Hall Effect technique. For VLSI, the challenge in annealing is not simply to repair damage and activate dopant, but to do so while minimizing diffusion so that shallow implants remain shallow. This has motivated much work in rapid thermal annealing (RTA), where annealing times are on the order of seconds. RTA uses tungsten-halogen lamps or graphite resistive strips to heat the wafer from one or both sides as against conventional furnace annealing where times or on the order of minutes.

Modern device structures, such as the lightly-doped drains (LDD) for MOSFET, require precise control of dopant distribution vertically and lateral on a very fine scale. For VLSI CMOS structure, we need to form shallow n and layers with implantation energies within the reach of standard machines. As stated earlier, the ion velocity, perpendicular to the surface, determines the projected range of an implanted ion distribution. If the water is tilted at a large angle to the ion beam then the effective ion energy is greatly reduced tilted ion beams, thus, make it possible to achieve extremely shallow dopant distributions using comparatively high implantation energies. We can circumvent the problem of implanting a shallow layer in silicon completely if instead we implant entirely into a surface layer and then diffuse the dopant into the substrate. This is most often done when the surface film is to be used as a conductor making contact to the substrate. Diffusion results in steep dopant profiles without damage to the silicon lattice.

Dopant diffusion in silicides and polysilicon is generally much faster than in single-crystal silicon, so the implanted atoms soon become uniformly distributed in the film.

4.3.6.7 Importance of Ion Implantation for VLSI Technology

Ion implantation is a very popular process for VLSI because it provides more precise control of dopants (as compared to diffusion). With the reduction of device sizes to the submicron range, the electrical activation of ion-implanted species relies on a rapid thermal annealing technique, resulting in as little movement of impurity atoms as possible. Thus, diffusion process has become less important than methods for introducing impurity atoms into silicon for forming very shallow junctions, an important feature of VLSI circuits. Ion, implantation permits introduction of the dopant in silicon that is controllable, reproducible and free from undesirable side effects. Over the past few years, ion implantation has been developed into a very powerful tool for IC fabrication. Its attributes of controllability and reproducibility make it a very versatile tool, able to follow the trends to finer-scale devices. Ion implantation continues to find new applications in VLS technologies.

4.3.7 Metallization

4.3.7.1 Metallization basics

Integrated circuit fabrication is traditionally divided into two segments, that follow one after the other in the fab.

1. **FEOL** (Front end of the line) - these refer to the fabrication of the active and passive elements of the circuit. These are the resistors (or conductors), capacitors, diodes, and transistors that make up the var- ious elements of the IC.

2. **BEOL** (Back end of the line) - these are the metallic layers that are used to make the interconnections between the various components fabricated in FEOL and also to the connections for the external devices.

With increase in device complexity, the separation of the IC processing into two segments is also important in terms of device fabrication. Current metal- lization in the IC industry is based on copper, which is a deep defect forming impurity in Si. Thus, Cu contamination in Si can destroy device functional- ity. By separating the fabrication into two segments, it is possible to isolate the Si processing from the

metals (primarily Cu) and prevent contamina- tion. There are strict *process and physical separation* between the FEOL and BEOL.**Metallization** refers to the "wiring" of the various components together to get a functioning circuit. In the first IC fabricated (by Jack Kilby) metal con- nections were made by external wiring (aluminum). Future devices, starting from the modifications made by Robert Noyce, had metal lines that were fabricated along with the IC. Typical steps in patterning a metal layer are shown in 4.3.7.1. There are a variety of techniques for depositing metal layers in a IC.

1. **Sputtering** is a physical vapor deposition process mainly used for Al and its alloys e.g. Al-Cu alloys

2. **Chemical vapor deposition** (CVD) is mainly used for poly Si (for gate in MOSFET) and tungsten (metal plugs for trench filling). It is also used for depositing barrier layers (silicides and nitrides) between Si and Cu.

3. **Electroplating** is used for Cu deposition (dual-damascene process)

With increase in level of integration, the metallization materials have changed. At the same time, the number of metal layers required have also increased (due to decrease in available area between the components). In MSI (medium scale integration), a single layer of metallization was sufficient, as shown in figure 4.3.7.1. But with increase in integration level, the number of metal layers have also increased.
A two level metallization scheme is shown in figure 4.3.7.2. The first level of metals provides the connection to the semiconductor i.e. the source, drain, and gates of a transistor. This is done by creating contact holes, using a photomask, a process called *contact masking*. Then, metal lines are vapor deposited and the excess metal is removed during lift-off. Usually, there is also a post an- nealing step for alloying. The metal lines are then further connected to each other, to form circuits, and then to the external devices by using a second

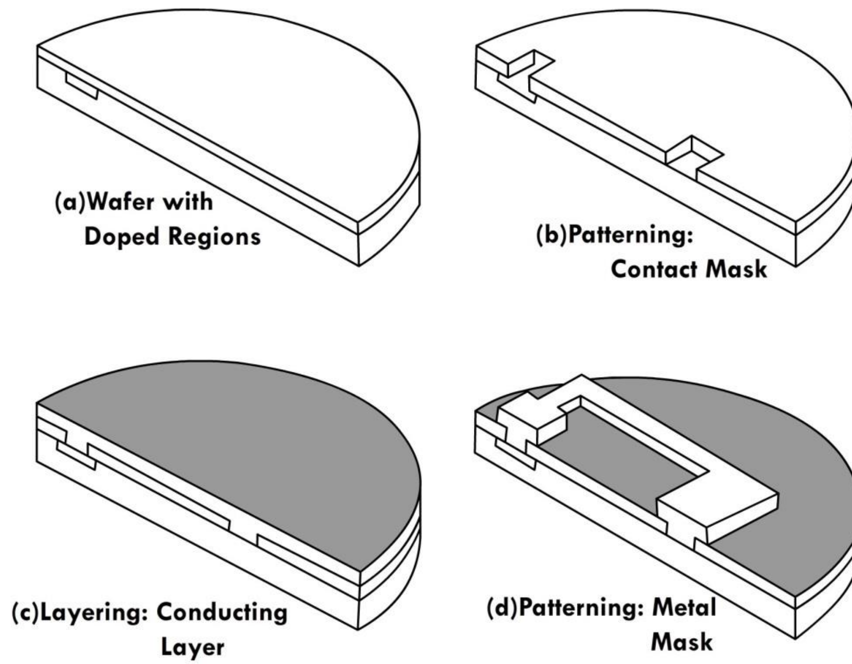


Figure 4.3.7.1: Sequence of steps in metallization. Here, the metal line connects a doped region to the rest of the wafer. For this, (a) the wafer is (b) patterned using a soft lithography mask. (c) The metal layer is deposited uniformly and (d) the mask, with the rest of the metal is removed.

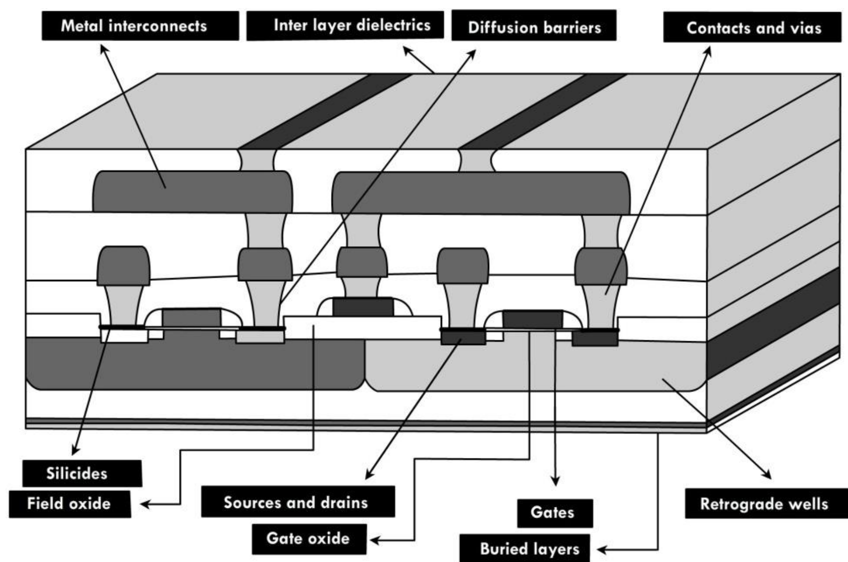


Figure 4.3.7.2: A two level metallization scheme.

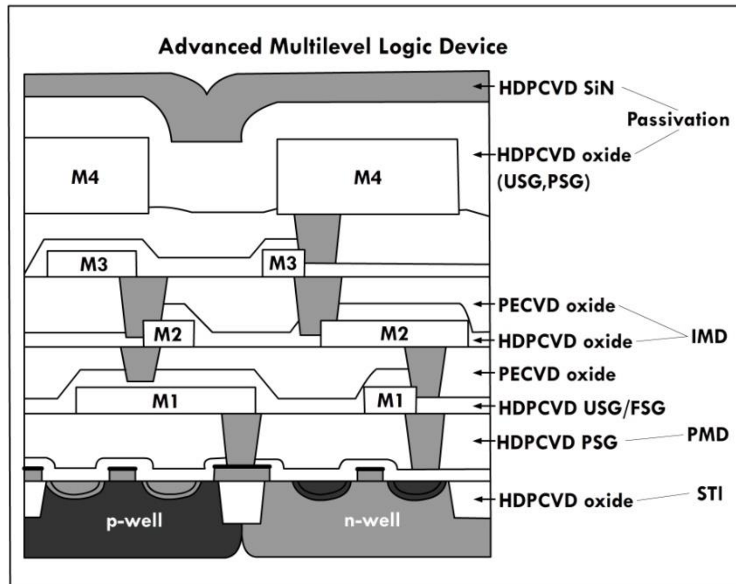


Figure 4.3.7.3: A four level metallization scheme.

level of metallization. The two levels are separated by interlayer dielectrics to prevent shorting. This is called *intermetallic dielectric layer* (IML). The levels can be extended to more than two, depending on the integration level. A four layer scheme is shown in figure 4.3.7.3. Current IC technology (28 *nm* technology) has **eleven** layers of metallization. A cross sectional image of the metal layers is shown in figure 4.3.7.4.

4.3.7.2 Metallization materials

Aluminum

The original metal used for wiring was pure Al. In the first circuit design proposed by Robert Noyce, pure Al was used for fabricating the wires. The main advantage of using Al is that it can be easily vapor deposited (simple thermal evaporation will work since Al has a low melting point). It also has good adhesion to SiO_2, low contact resistance, and it is easy to pattern since

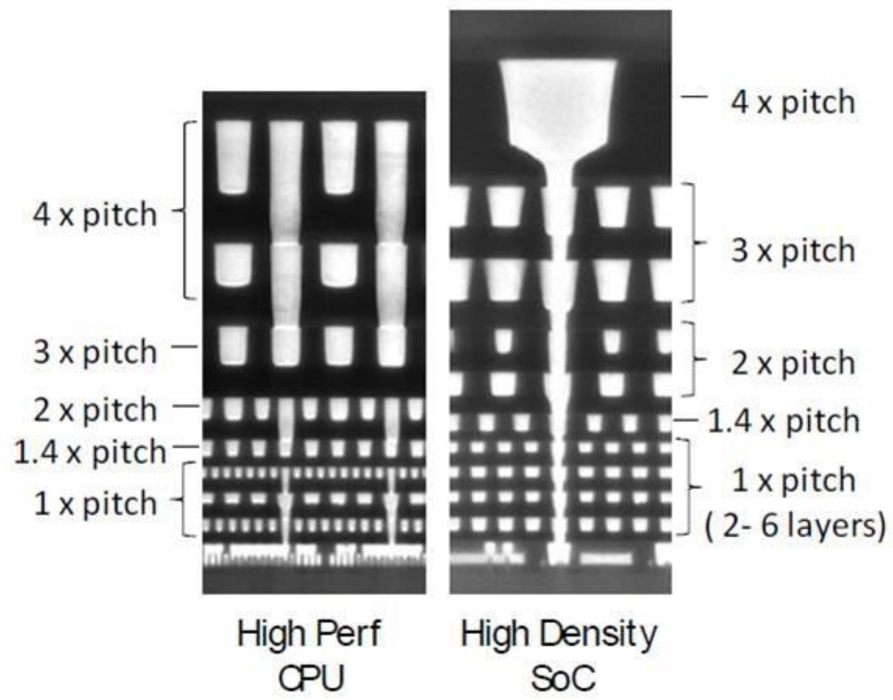


Figure 4.3.7.4: Eleven layers of metallization.

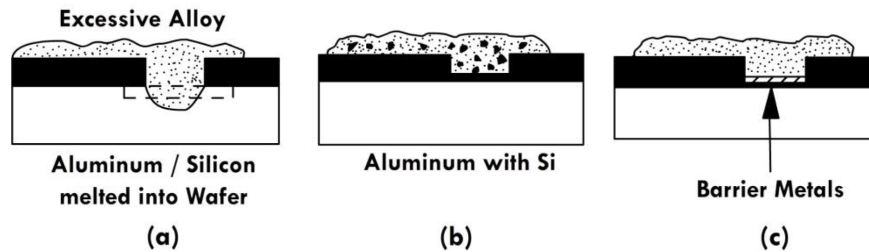


Figure 4.3.7.5: Contact issues in Al-Si contacts. (a) Excess alloying leads to melting of the Al(b) Silicide formation in the metal layer, by using a Al-Si alloy (c) Barrier metal is usually deposited to prevent reaction between Al and Si. thermal evaporation can be integrated with resist lithography technology.

Al-Si alloys

The problem with pure Al is that it has a low melting point of 660 ˚C. When Al in contact with pure Si is heated, it forms an alloy with an eutectic point of 577 C. This leads to dissolution of metal, especially in the formation of shallow junctions, and can lead to shortening of the contacts, as shown in figure 4.3.7.5.There are two solutions to this. One is to use a barrier metal that does not alloy with Al or Si and separates the two. The barrier metal should not significantly reduce the conduction through the channel. Typically, high temperature metals like Ti and W or compounds like TiN are used. These are sputter deposited on the wafer. Another option is to use Al with 1-2% Si as the contact material. This minimizes Al alloying with the Si wafer but does not eliminate it completely. Thermal evaporation of Al-Si might now work due to the large difference in the melting points of the two elements and other techniques like sputtering or e-beam evaporation are needed to maintain compositions of the contact.

Al-Cu alloys

With increase in device integration (from MSI to LSI and VLSI), the thick- ness of the metal layers decreases. This leads to the problem of **electro- migration**, especially in thin Al layers. This is because thin films with an electrical field gradient, due to the applied voltage, also develop a thermal gradient due to

resistive heating. The thermal gradient is acute for thinner layers since their resistance is higher. This causes local heating and migration of material from thinner areas of the wire, which can cause an open circuit. To reduce electromigration, 0.5-4% Cu is usually added to Al. Cu alloys with Al, to form $CuAl_2$ precipitates (GP zones). These precipitates pin the grain boundaries and reduce electromigration. Sometimes Si is also added to prevent Si dissolution from the wafer. The typical alloy composition for a metal layer is Al-1.5%Si-4%Cu.

Pure Cu

With smaller metal layers, Al-Cu has a high resistance (high resistivity of Al alloy) and hence to increase wire conductance pure Cu replaced Al as the metallization layer. Pure Cu contacts were introduced by IBM in 1990s and the standard was quickly adopted across the industry. Cu can be easily met- allized. It can be deposited by thermal evaporation, but more importantly, it can be electroplated on the wafer, which decreases the cost, since expensive vacuum chamber equipment is not needed. The biggest problem is that Cu diffuses into Si and SiO_2. These form deep level defects in Si which can *'kill'* the device. Hence, a barrier metal, usually TiW or TiN or TaN or metal sili- cides, is needed. These can be deposited by sputtering or for deep trenches, can be deposited by chemical vapor deposition. As mentioned earlier, the use of Cu separates the wafer manufacturing into FEOL and BEOL, with strict physical separation between the two to prevent contamination. Usu- ally, equipment involved in FEOL and BEOL are placed in different locations in the fab and special clothing is used for people working with BEOL tools.

4.3.7.3 Metallization techniques

4.3.7.3.1 Physical vapor deposition (PVD)

There are a variety of physical vapor deposition techniques. As the name implies, atoms/molecules (vapor) of the desired material are directly de- posited on to the substrate from the vapor phase. There are different PVD techniques, which differ on the how the 'vapor' is obtained. PVD is a *line- of-sight* deposition technique, so that the substrate must be in front of the source. The deposition rate depends on the distance between the two. The simplest PVD technique is *thermal evaporation*. A schematic of the process is shown in figure 4.3.7.6. The material to be evaporated is heated (by resistive heating) and the atoms are then deposited on

the substrate.*E-beam evaporation*, is a deposition technique, where instead of using resistive heating to form the vapor, an electron beam is used to melt the material and form the vapor.

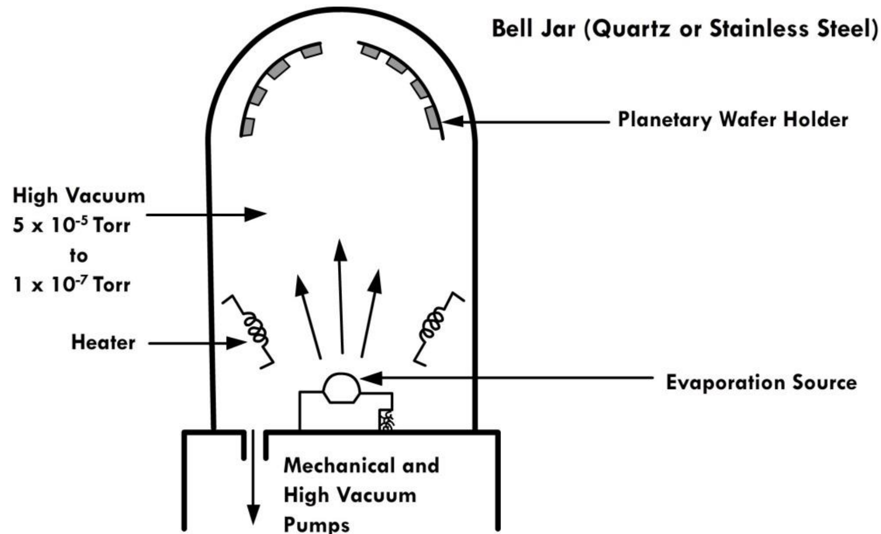


Figure 4.3.7.6: Thermal evaporator unit.

The e-beam evaporation source is shown in 7. E-beam evaporator is useful for depositing materials with high melting points like Si, Ti, W, which cannot be easily deposited by thermal evaporation.

Both, thermal evaporation and e-beam evaporation have deposition rates of a few $\AA$ per second. For depositing thick films (few hundred *nm* to μm), *sputtering* is used. The schematic of the sputter deposition process is shown in 4.3.7. 8. In sputter deposition, the material to be deposited is made the target electrode. This can be a pure metal, alloy or even compounds. Sput- tering process can maintain the stoichiometry of the target electrode, unlike thermal or e-beam evaporation. An inert gas, like argon, is introduced in the vacuum chamber. They are ionized by using an electron beam and the accel- erated ions strike the target electrode and remove material, a process called *sputtering*. Thus, the 'vapor' is created by the positively charged ions. The vapor atoms are then deposited on the substrate. The advantage of sputter- ing is that deposition rates of a few *nm* per second can be easily obtained. There are three main sputtering techniques: DC, RF and magnetron sput- tering. Their difference lies in how the Ar ions are accelerated and made to strike the target. In magnetron sputtering, magnetic fields are used to confine the electrons in front of the target to increase the ionization of the Ar gas and thus

increase deposition rate.

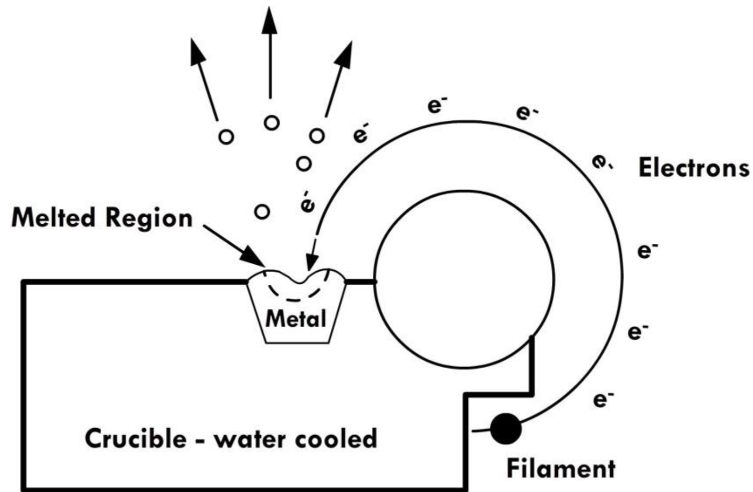


Figure 4.3.7.7: An e-beam evaporator unit.

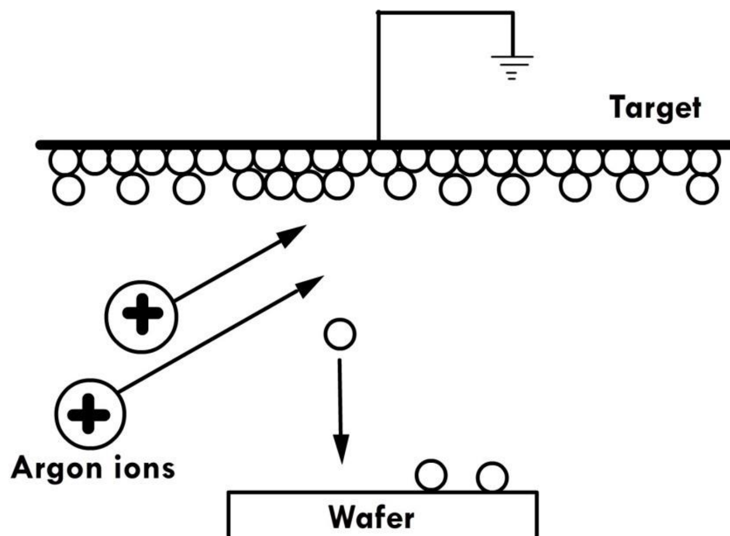


Figure 4.3.7.8: Schematic of the sputter deposition process.

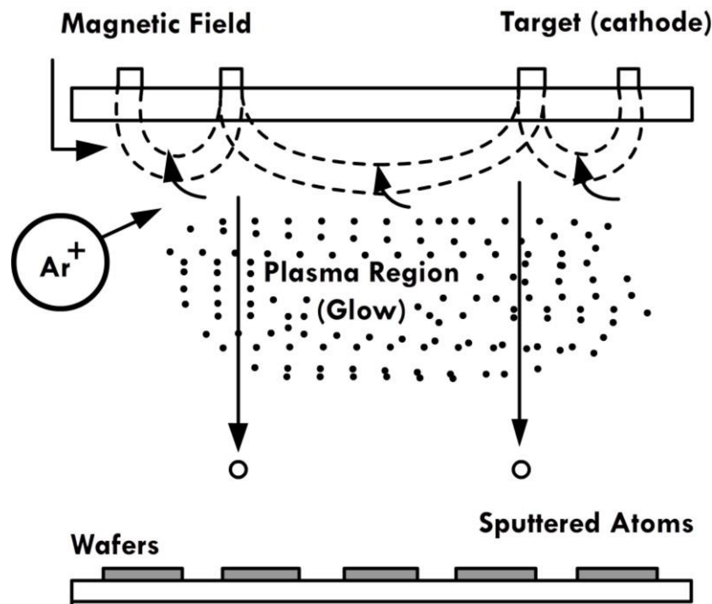
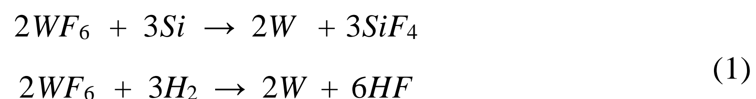


Figure 4.3.7.9: Schematic of the magnetron sputtering process.

This also results in lower chamber pressure requirement, making this a cleaner process. The schematic of the magnetron sputtering process is shown in figure 4.3.7.9.

4.3.7.3.2 CVD

The CVD process was seen earlier in the context of deposited films. In metallization, CVD process is used for the deposition of the barrier layer that separates the metal from Si. This is good for *large aspect ratio structures*, as shown in figure 4.3.7.10. A typical reaction for depositing tungsten is by the reduction of tungsten hexaflouride.



$$2WF_6 + 3Si \rightarrow 2W + 3SiF_4$$
$$2WF_6 + 3H_2 \rightarrow 2W + 6HF \quad (1)$$

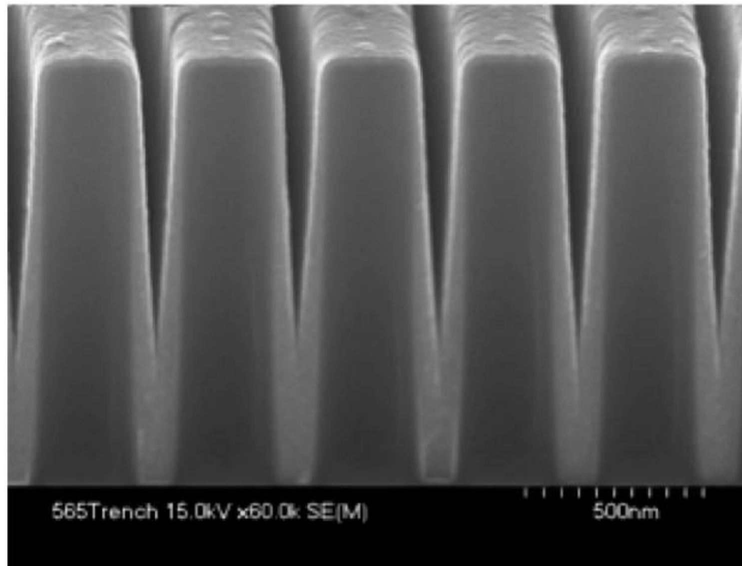


Figure 4.3.7.10: CVD process for growing conformal layers on deep trenches. Deposition on large aspect ratio structures cannot be done by PVD techniques, since the method will cover the hole before depositing deep inside the trench.

4.3.7.3.2 Electroplating

The electroplating process is commonly used for the deposition of copper. The advantage is the low cost and temperature requirements, compared to other vacuum deposition techniques. High deposition rates can be obtained, compared to PVD processes. Electroplating requires a uniform seed layer. This is obtained by sputtering and the seed layer is 30-200 *nm* thick. The electroplating bath is shown in figure 4.3.7.11. The wafer, containing the seed layer, is made the cathode. The copper to be deposited is the electrolyte (in the form of $CuSO_4$). This is reduced in the bath (equation shown in figure 4.3.7.11) and the Cu is then plated on the wafer surface. The process leads to
a copper overfill, creating a rough surface and the excess material is then removed by polishing.

4.3.7.4 Planarization

Planarization is a process of achieving a flat profile on the wafer surface. This is important for lithography, since a flat wafer is needed to avoid registration errors when the mask is aligned with the wafer and to get proper focusing.

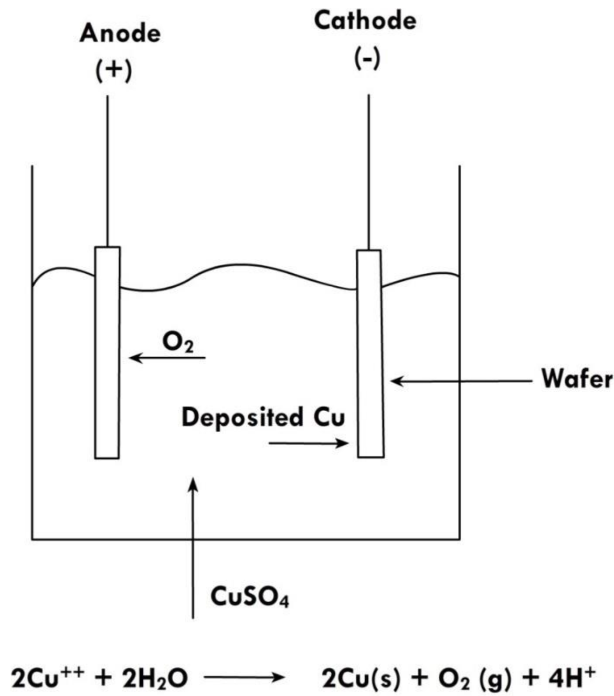


Figure 4.3.7.11: Schematic of the electroplating process.

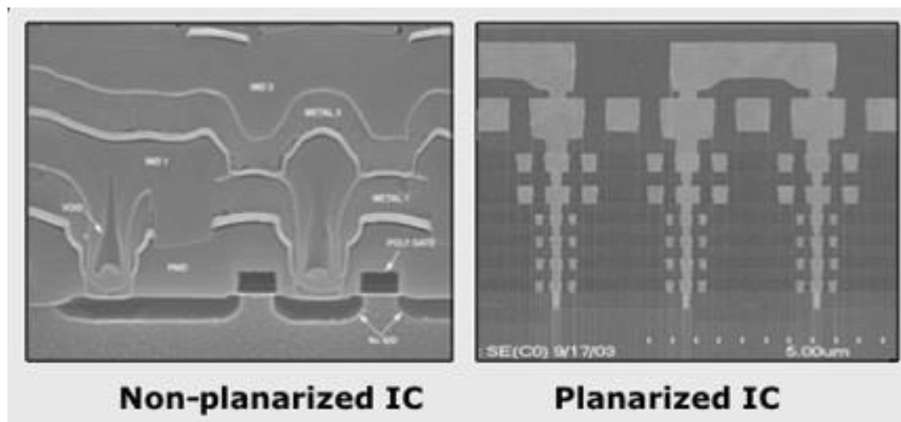


Figure 4.3.7.12: Non planarized vs. planarized IC.

Most deposition process produce a surface with a finite rough- ness which usually increases with increased deposition rate and thickness. **Chemical mechanical polishing (CMP)** is a technique to achieve global planarization, i.e. over the entire wafer. The difference between a planarized and non-planarized wafer is shown in figure 4.3.7.12.

In planarization, the wafer is mounted on a rotating platen. It is then pol- ished using a polishing pad and a slurry containing abrasive particles. The abrasive particles attack the wafer surface and remove small particles. The polishing pad and platen rotate in opposite directions and the slurry carries away the small particles. This removal is the mechanical polishing part. The slurry material is chosen such that it can also dissolve or etch the surface material away. This constitutes the chemical removal part and hence, the technique is called CMP. The schematic of the CMP process is shown in fig- ure 13. Typically, the polishing pad is made of polyurethane foam, while the slurry depends on the material to be removed. For metals, usually alumina is used, while etchants like KOH and NH_4OH are used for silicon oxide polishing. After CMP, there is a post cleaning step that involves cleaning thewafers with de-ionized water and then N_2 blow drying. This removes any excess slurry particles from the wafer surface.

4.3.7.5 Packaging

4.3.7.5.1 Introduction

The various components of the IC are connected to each other through the metallization layers. There are various levels in metallization, which provide different levels of connectivity (the latest Intel processors have 11 levels of metallization). The final metallization layer is used to connect the IC to external devices. Packaging refers to the set *of processes* that provide this electrical connectivity, thus allowing the chip to be integrated with other devices. Packaging is also used to provide *physical protection* to the chip. For a given system, the various components can be integrated into one chip. This is called *System on Chip (SoC)*. There are also situations where multiple chips with specific functionality are integrated on a package, called *System on Package* (SoP).

There are various chip characteristics that affect the packaging process

1. Integration level
2. Wafer thickness

3. Chip dimensions
4. Environmental sensitivity - important for lead free packaging
5. Physical vulnerability
6. Heat generation
7. Heat sensitivity - heat generation and sensitivity are important during operation since adequate provision must be given for heat dissipation.

Usually, a passivation layer is grown on top of the wafer, at the end of the fabrication. This is provide for protection to the circuit elements. This passivation layer can be a hard layer (like silicon nitride or oxide) or a soft layer like polyimide. The electrical contacts are exposed, so that the chip can be connected to the system. There are four basic functions of apackage

1. Substantial lead system - electrical connectivity
2. Physical protection
3. Environmental protection
4. Heat dissipation - this is very important for chips that are used in mobile computing, where active cooling elements cannot be implemented. Lower heat dissipation, by reduced power consumption, is also impor- tant for mobile computing since this save battery life.

Packaging is done after the wafers are done with the fabrication and sort process. The sort process isolates the dies that are good and need to be packaged. Packaging is a series of steps, similar to the assembly line process in fabrication, through which the dies pass before final inspection and delivery.

4.3.7.5.2 Packaging process

Backside prep

This is the first step in packaging, where the wafers are thinned. Typical 12" wafers, used in fabrication, have a thickness of 650-700 μm. This thickness is needed for wafer handling during fabrication, but less than 100 *mu*m of Si is used. The thickness of the wafers can affect die separation and packaging and hence the wafers need to be thinned to approximately 100 *mu*m. This is done by chemical mechanical polishing. The front side is protected during this process (using thick resist) to make sure no cracks or defects are introduced during thinning.

Die separation

The wafers are separated into individual dies. This is usually done by sawing or scribing the wafer, using a diamond saw or scribe, on pre patterned scribe lines. These scribe lines also have e-test structures, that are probed during sort, to identify the die yield. Scribe lines in a wafer are shown in figure 4.3.7.13.

Die pick and place

After die separation, the good dies are picked up for further processing. Prior to this, good/bad dies are identified in sort. The die sort, separation and pick process is shown in figure 4.3.7.14

Die inspection

After die pick, the dies are inspected for cracks/defects on the good dies. This inspection is done using an optical microscope and the process is automated.

Die attach

After passing inspection, the die is then attached to the specific area of the package. An example is shown in figure 4.3.7.15. The die attach process is used to create a strong bond between the die and package and is the precursor for the

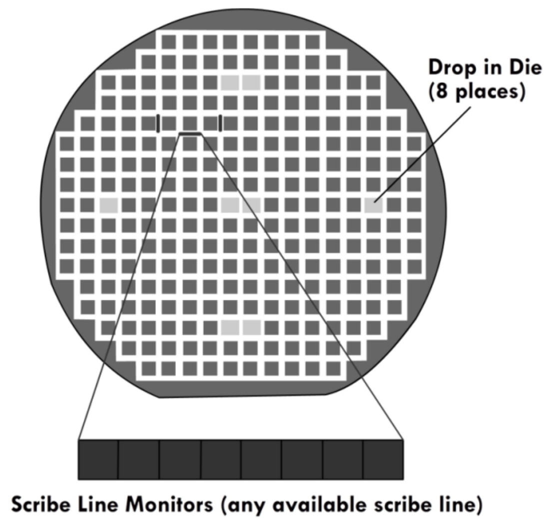


Figure 4.3.7.13: Scribe lines in a wafer separating individual dies.

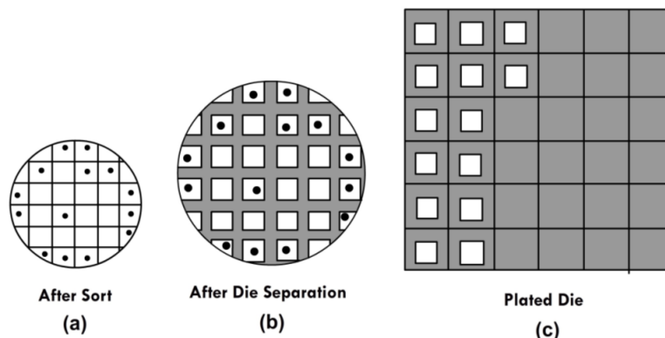


Figure 4.3.7.14: (a) A die sort process to identify good dies, (b) Separation of good dies, and (c) picking the good dies for further process. Bad dies are

usually scrapped and the silicon reused.

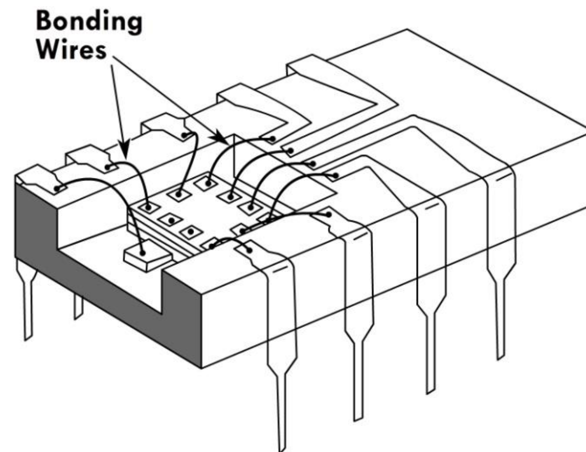


Figure 4.3.7.15: Schematic of a die attached to the package.

wire and tape bonding process, during which electrical leads are attached. For the flip chip process, the die attach and bonding process are combined. The die attachment can be conductive e.g. using a gold-silicon eutectic. The eutectic has a melting point of 380 ˚C, shown in the phase diagram in figure 4.3.7.16. Hence, gold is plated on the bottom of the die, in contact with Si and then heated, so that the eutectic is formed and melts to form the bond. Non conducting attachments are also used. In this case, a epoxy adhesive material is used to attach the die to the package.

Bonding

Bonding is the most important step in the packaging process. In bonding, wires are connected to the leads in the IC, so that the IC is *electrically connected* with other devices. Bonding usually follows die attach, shown in figure 4.3.7.15 or the two steps are combined, as in the flip chip process. There are three main techniques for bonding

> **Wire bonding** - this is carried out using gold or aluminum. The gold wire is fed through a capillary and by thermo mechanical compression is bonded with the die and lead. The schematic of this process is shown in figure 4.3.7.17.This type of bonding is also called *ball bonding*. In the case of Al, a wedge bond is formed between the die and the lead. This is

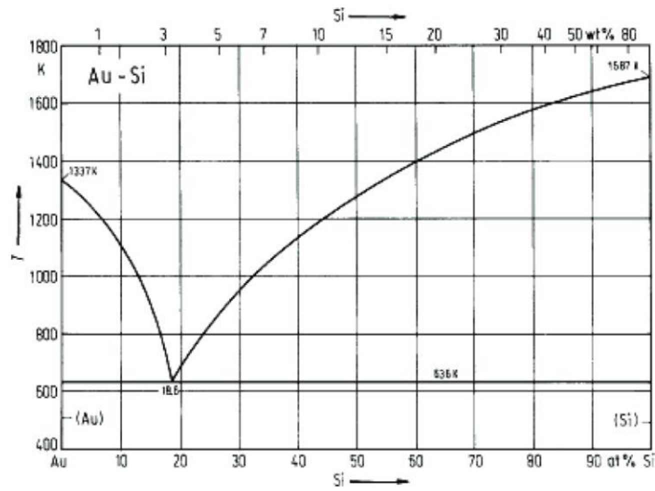


Figure 4.3.7.16: Au-Si phase diagram, with a deep eutectic point at 380 °C.

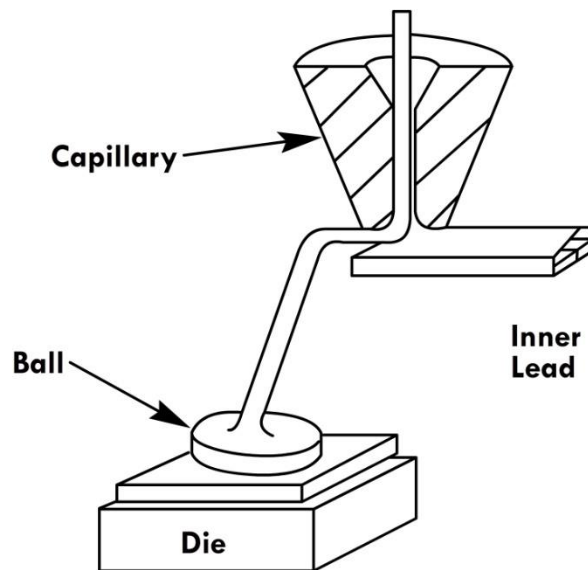


Figure 4.3.7.17: A gold wire bonding to the lead and die using a ball bonding.

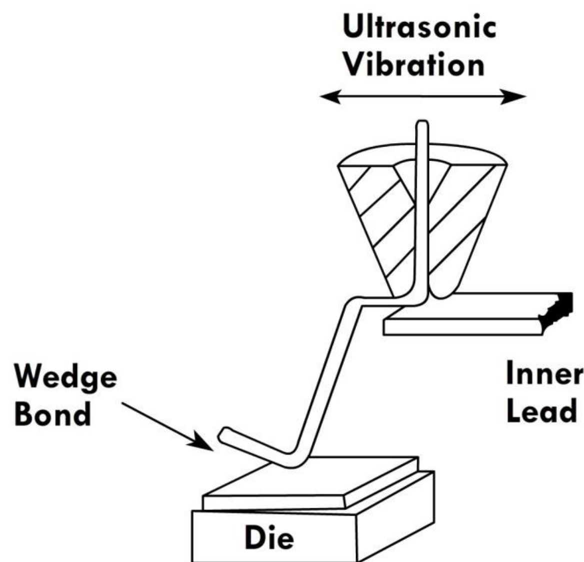


Figure 4.3.7.18: An aluminum wedge bonding between the lead and die.

achieved by using ultrasonic vibration. The advantage of using Al is that, lower temperatures are needed and it is cheaper than gold. The Al wedge bond is shown in figure 4.3.7.18.

Tape bonding - tape bonding is used for extremely thin device fabrication. It is also called *tape automated bonding* (TAB). In this process, the electrical lead system is deposited on a flexible tape by sputtering or thermal evaporation.This is combined with a patterning or stamping process to make the leads on the tape. For bonding, the die is position and aligned with the leads on the tape and the bonding is completed by a tool called the *thermode*. The thermode has a heated flat diamond surface that forces the tape onto the die and makes the bond. The process is shown in figure 4.3.7.19

Flip-chip bonding - this is also called *bump and ball bonding* or *controlled collapse chip connection* (C4). It is currently in use in the IC industry, due to the large number of leads and the small spacing between them. In this process, the wires are replaced by a reflowed solder bump, as shown in figure 4.3.7.20. The metal bumps are deposited on top of the leads on the die and the chip is then flipped

(hence the name flip chip) and bonded with the package using the solder bumps.

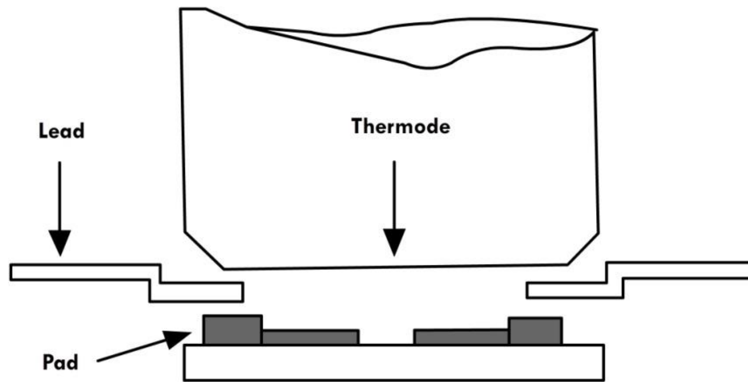


Figure 4.3.7.19: Schematic of the tape bonding process.

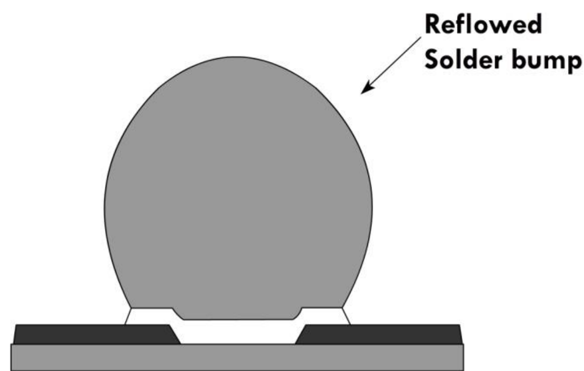


Figure 4.3.7.20: Reflowed solder bump that is used for flip-chip bonding.

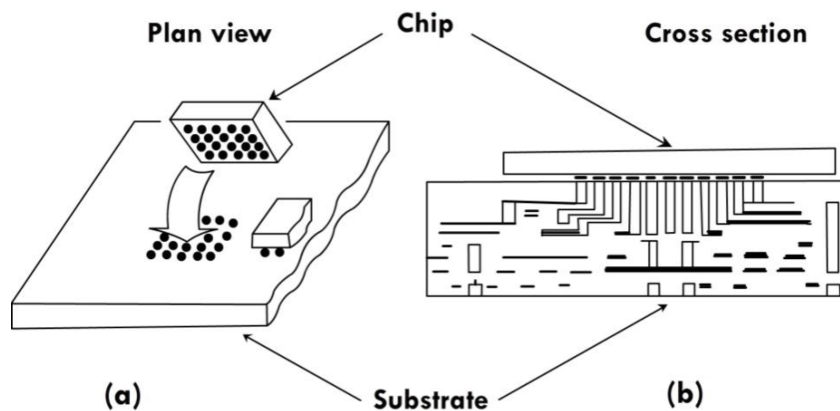


Figure 4.3.7.21: Schematic of the Flip-chip or C4 bonding process in (a) plan view and (b) cross section.

The process is shown in figure 4.3.7.21.There is no separate die attach step, since both attach and bonding takes place through the solder bumps. For additional mechanical stability, an epoxy filling is also used. The formation of the reflow solder process takes place partly in the fab, through a series of processing steps. These steps are summarized in figure 4.3.7.22. Some of the steps are carried out during fabrication, in the back end process.

Pre seal inspection

After bonding, there is an inspection step to check the bonds formed. This inspection uses optical measurements, to check the bonding and reliability measurements for checking the electrical connectivity. There are different reliability criteria depending on the nature of the application.

Sealing

After inspection, the die is sealed in a protective enclosure. Seals can be her- metic or non-hermetic, depending on the application. Hermetic seals help in isolating the die from the atmosphere and include welding,

soldering, and glass sealing. Non hermetic seals use epoxy molding, while hermetic seals are usually premade packages, either with ceramic or low melting glass (CERDIP). The parts of the CERDIP package are shown in figure 4.3.7.23.

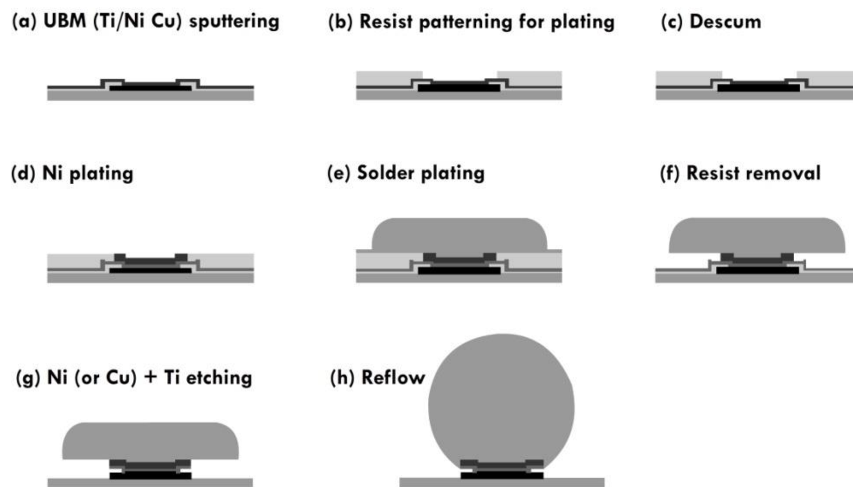


Figure 4.3.7.22: (a)-(h) Sequence of steps in the fabrication of the reflow solder over the die leads. Some of these steps are carried out during IC fabrication. They involve patterning the leads and depositing suitable conductive barrier materials. The solder is then deposited and heated to assume the shape of the reflow.

Final steps

Once the die is sealed, the leads are coated with lead-tin solder. This is to help in bonding the package to a printed circuit board. It also helps in proving the electrical connectivity with other devices. This process is called *lead plating*. Then the excess lead is removed, a process called *lead trimming*. Then, the package is marked, *package marking*, and then the final step is electrical testing, appropriately called *final testing*. Packages that pass final testing are then sent for assembly or shipped to the customer.

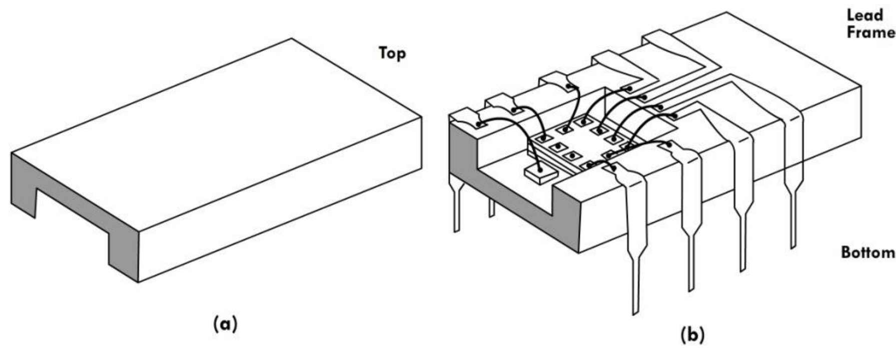
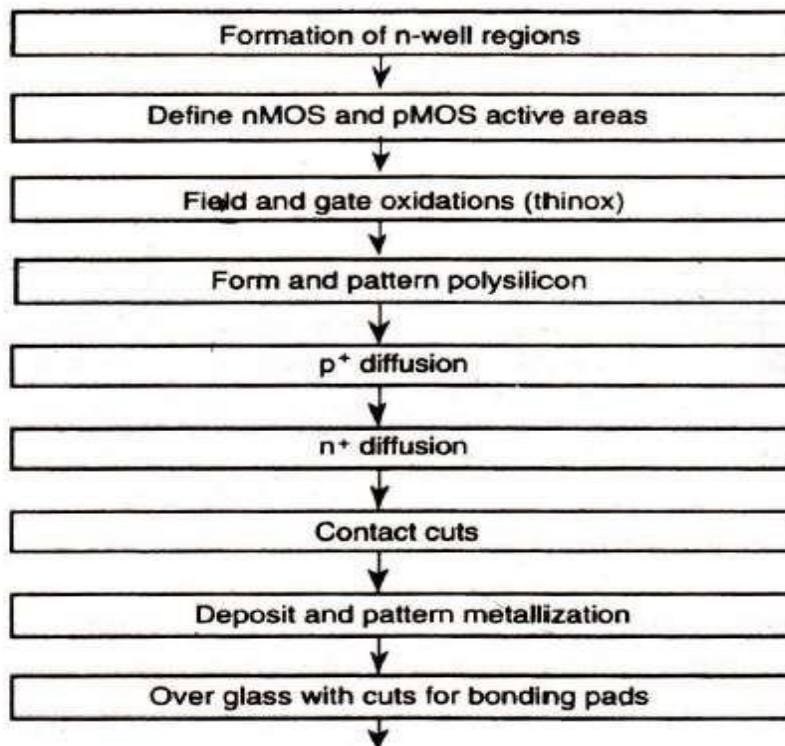


Figure 4.3.7.23: (a) Top and (b) bottom part of the CERDIP package. The top part has a cavity where the die sits and it is coated with low melting glass to form a hermetic seal. The bottom part contains, the die with all the electrical leads attached.

4.3.8 CMOS n Well Process

The n-well Process : Though the p-well process is widely used in C-MOS fabrication the n-well fabrication is also very popular because of the lower substrate bias effects on transistor threshold voltage and also lower parasitic capacitances associated with source and drain regions.

The typical n-well fabrication steps are shown in the diagram below.



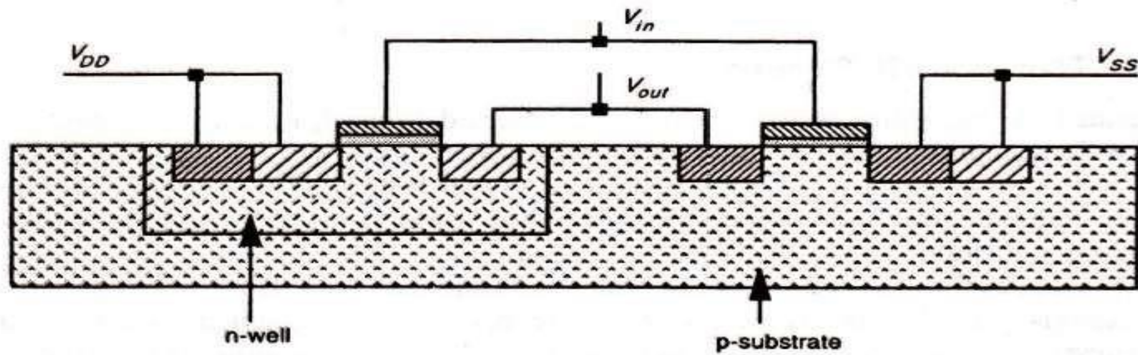


Figure 4.3.8.1. n-well CMOS Inverter.

The first mask defines the n-well regions. This is followed by a low dose phosphorus implant driven in by a high temperature diffusion step to form the n-wells. The well depth is optimized to ensure against-substrate top+ diffusion breakdown without compromising then-wellton+ mask separation. The next steps are to define the devices and diffusion paths, grow field oxide, deposit and pattern the polysilicon, carry out the diffusions, make contact cuts, and finally metalize as before. lt will be seen that an n+ mask and its complement may be used to define the n- and p-diffusion regions respectively. These same masks also include the VDD and Vss contacts (respectively). It should be noted that, alternatively, we could have used a p+ mask and its complement since the n + and p + masks are generally complementary. Due to the differences in charge carrier mobilities, the n-well process creates non-optimum p- channel characteristics. However, in many CMOS designs (such as domino-logic and dynamic logic structures), this is relatively unimportant since they contain a preponderance of n-channel devices. Thus then-channel transistors are mainly those used to form1ogic elements, providing speed and high density of elements.However, a factor of the n-well process is that the performance of the already poorly performing p-transistor is even further degraded. Modern process lines have come to grips with these problems, and good device performance may be achieved for both p-well and n-well fabrication.

4.3.8.2 The p-well Process :

The p-well structure consists of an n-type substrate in which p-devices may be formed by suitable masking and diffusion and, in order to accommodate n-type devices, a deep p-well is diffused into the n-type substrate as shown in the Fig.below. This diffusion should be carried out with special care since the p-well doping concentration and depth will affect the threshold voltages as well as the breakdown voltages of the n-transistors. To achieve low threshold voltages (0.6 to 1.0 V) either deep-well diffusion or high-well resistivityis required. However, deep wells require larger spacing between the n- and p-type transistors and wires due to lateral diffusion and therefore a larger chip area. The p-wells.act as substrates for the n- devices within the parent n-substrate, and, the two areas are electrically isolated.Except this in all other respects- like masking, patterning, and diffusion-the process is similar to nMOS fabrication.

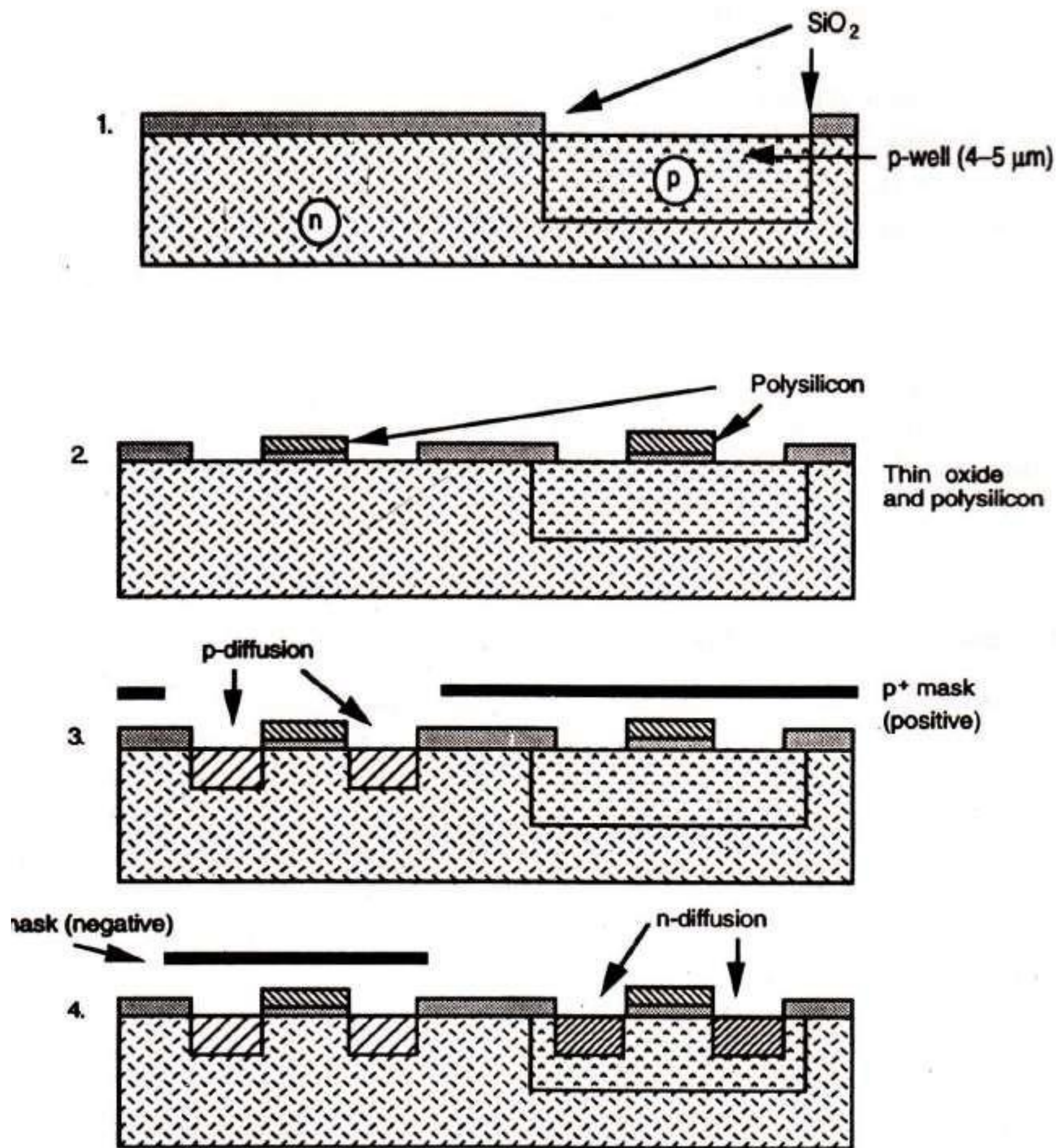
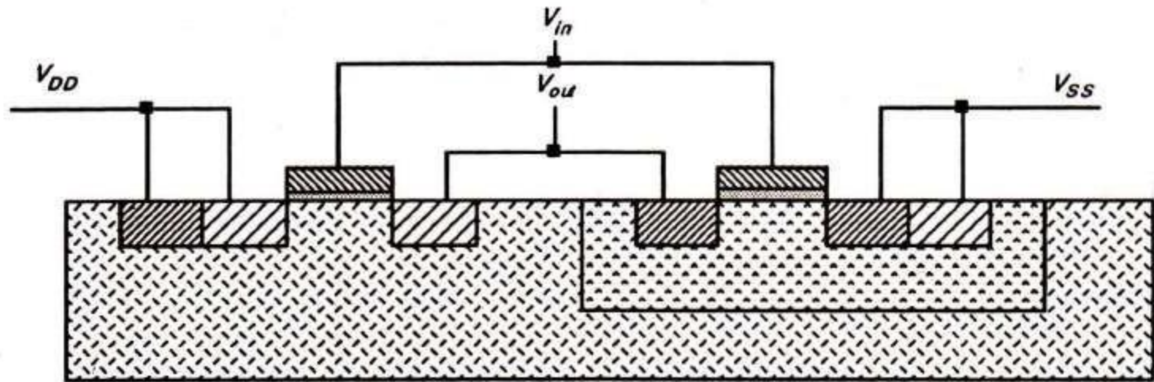
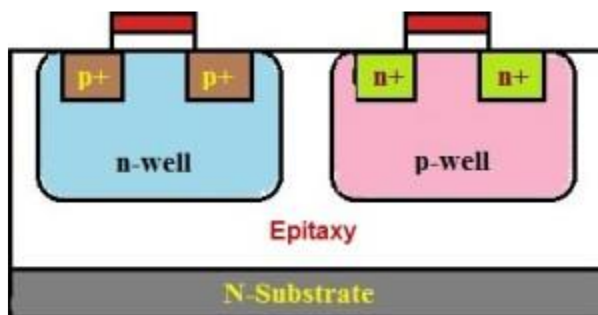


Figure 4.3.8.2 p-well fabrication process(Figs 1,2,3 & 4)

The diagram below shows the CMOS p-well inverter showing VDD and Vss substrate connections.



4.3.8.3 Twin tub-CMOS Fabrication Process



In this process, separate optimization of the ntype and p type transistors will be provided. The independent optimization of Vt, body effect and gain of the P-devices, N-devices can be made possible with this process.Different steps of the fabrication of the CMOS using the twin tub process are as follows:

- Lightly doped n+ or p+ substrate is taken and, to protect the latch up, epitaxial layer is used.
- The high-purity controlled thickness of the layers of silicon are grown with exact dopant concentrations.
- The dopant and its concentration in Silicon are used to determine electrical properties.
- Formation of the tub
- Thin oxide construction
- Implantation of the source and drain
- Cuts for making contacts
- Metallization

By using the above steps we can fabricate CMOS using twin tub process method.In Duel-well process both p-well and n-well for NMOS and PMOS transistors respectively are formed on the same substrate. The main advantage of this process is that the threshold voltage, body effect parameter and the transconductance can be optimized separately. The starting material

for this process is p+ substrate with epitaxially grown p-layer which is also called as epilayer. The process steps of twin-tub process are shown in Figure below.

The process starts with a p-substrate surfaced with a lightly doped p-epitaxial layer.

Step 1 : A thin layer of SiO_2 is deposited which will serve as the pad oxide.

Step 2 : A thicker sacrificial silicon nitride layer is deposited by chemical vapour deposition.

Step 3 : A plasma etching process is used to create trenches used for insulating the devices.

Step 4 : The trenches are filled with SiO_2 which is called as the field oxide.

Step 5 : To provide flat surface chemical mechanical planarization is performed and also sacrificial nitride and pad oxide is removed.

Step 6 : The p-well mask is used to expose only the p-well areas, after this implant and annealing sequence is applied to adjust the well doping. This is followed by second implant step to adjust the threshold NMOS transistor.

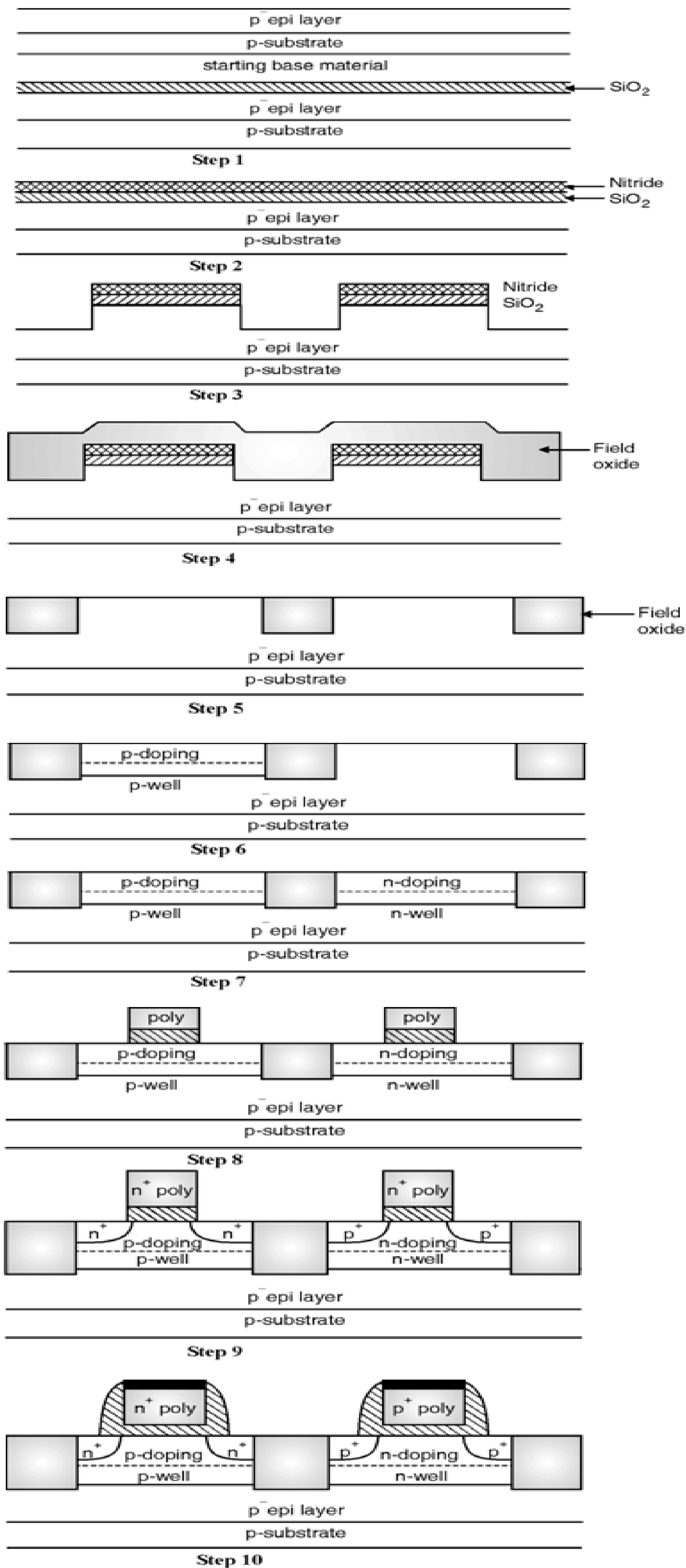
Step 7 : The n-well mask is used to expose only the n-well areas, after this implant and annealing sequence is applied to adjust the well doping. This is followed by a second implant step to adjust the threshold voltage of PMOS transistor.

Step 8 : A thin layer of gate oxide and polysilicon is chemically deposited and patterned with the help of polysilicon mask.

Step 9 : Ion implantation to dope the source and drain regions of the PMOS (p^+) and NMOS (n^+) transistors is used this will also form n^+ polysilicon gate and p^+ polysilicon gate for NMOS and PMOS transistors respectively.

Step 10 : Then the oxide or nitride spacers are formed by chemical vapour deposition (CVD).

Step 11 : In this step contact or holes are etched, metal is deposited and patterned. After the deposition of last metal layer final passivation or overglass is deposited for protection.


p⁻ epi layer
p-substrate
starting base material
SiO2
p⁻ epi layer
p-substrate
Step 1
Nitride
SiO2
p⁻ epi layer
p-substrate
Step 2
Nitride
SiO2
p⁻ epi layer
p-substrate
Step 3
Field oxide
p⁻ epi layer
p-substrate
Step 4
Field oxide
p⁻ epi layer
p-substrate
Step 5
p-doping
p-well
p⁻ epi layer
p-substrate
Step 6
p-doping
n-doping
p-well
n-well
p⁻ epi layer
p-substrate
Step 7
poly
poly
p-doping
n-doping
p-well
n-well
p⁻ epi layer
p-substrate
Step 8
n+ poly
n+ poly
n+
n+
p+
p+
p-doping
n-doping
p-well
n-well
p⁻ epi layer
p-substrate
Step 9
n+ poly
p+ poly
n+
n+
p+
p+
p-doping
n-doping
p-well
n-well
p⁻ epi layer
p-substrate
Step 10

Sample Questions MODULE -IV

4.1 Multiple Choice Questions

4.1.1) Lambda based rule is

a) Dependent on feature size

b) Only for wires

c) Only for CMOS

d) all false

4.1.2) Stick encoding n-diffusion represnted by colour

a) yellow

b) green

c) red

d) blue

4.1.3) Diffusion rate of impurities into semiconductor lattice depends on

a) Mechanism of diffusion

b) Temperature

c) Physical properties of impurity

d) all true

4.1.4) In lambda(λ) based design rule minimum width of n-Diffusion is

a) 3 λ

b) 2 λ

c) λ

d) 6 λ

4.1.5) Fick's Law is associated with

a) Ion Implantation

b) Stick Encoding

c) Mask Design

d) Diffution

4.1.6) **Sputtering is a**

A) Physical Vapor Deposition process

b) Chemical Vapor Deposition process

c) Is a **Electroplating process**

d) all false

4.1.7) Physical vapor deposition is used for

a) Etching

b) Ion Implantation

c) Metallization

d) Diffution

4.1.8) Czochralski technique is used for

a) Metal deposition over silicon

b) Diffusion

c) Single crystal silicon manufacturing

d) All False

4.1.9) p-type(111) Si has

a) 2 primary flats

b) 1 primary flast

c) 3 primary flat

d) all false

4.1.10) Float zone technique used for

a) Single crystal silicon manufacturing

b) Silicon etching

c) Diffusion

d) Metallization

4.2 Short answer type questions

4.2.1) What is stick diagram? Draw the NAND gate stick diagram representation.

4.2.2) Briefly describe lambda based design rules.

4.2.3) State the lambda based design rules for Design rules for nMOS,pMOS and CMOS

4.2.4) What is meant by etching?Briefly describe etching process with diagrams

4.2.5) Draw the stick diagram and layout for NOR gate.

4.3 Long answer type questions

4.3.1)What is a twintub process? Describe the twin tub process with suitable diagram and elaborate each step. (3+12)

4.3.2) What are the different ways of fabrication of CMOS.Describe p-well and n-well process with process steps and suitable diagrams (3+12)

4.3.3) What is photolithography? How a design can be transferred to silicon using photolithography? What are the different types of photo-resists and state their role in photolithography? (2+8+5)

4.3.4) State the different types of diffusion mechanism. State Fick's law of diffusion .Briefly describe the wet and dry etching process. (3+2+10)

4.3.5) What is Ion Implantation? Draw the diagram of an ion implantation system and explain its operation. What are the advantages of ion-implantation?
(2+10+3)

Module-5

Introduction to Low Power and High Speed VLSI Circuit Design

5.1 Power consumption in CMOS inverter

Three types of power consumptions are observed in CMOS inverters a) dynamic power b) short circuit power and c) leakage power dissipations

Dynamic power in CMOS inverter is due to switching of out between high and low state . The expression of dynamic power as follows

$$P_{avg} = \alpha_T \cdot C_{load} \cdot V_{DD}^2 \cdot f_{CLK}$$

Whenever input signal is rising or falling the short circuit power dissipation occurs in the moment when both NMOS and MOS are in on condition . The expression of short circuit power dissipation as follows

$$I_{avg}(short - circuit) = \frac{1}{12} \cdot \frac{k \cdot \tau \cdot f_{CLK}}{V_{DD}} \left(V_{DD} - 2\,V_T\right)^3$$

Leakage power dissipations occurs due to reverse biasing of P-N junction at source to bulk and drain to bulk .

5.2 Timing parameters

Critical path : The critical path is defined as the path between an input and an output with the maximum delay

Arrival time : The arrival time of a signal is the time elapsed for a signal to arrive at a certain point

Slack: Slack is defined as difference between actual or achieved time and the desired time for a timing path.

Skew : Clock skew (sometimes called timing skew) is a phenomenon in synchronous digital circuit systems (such as computer systems) in which the same sourced clock signal arrives at different components at different times

Set-up time : Setup time is defined as the minimum amount of time before the clock's active edge by which the data must be stable for it to be latched correctly

Hold time : Hold time is defined as the minimum amount of time after the clock's active edge during which data must be stable

Gate delay and path delay: This is the delay , time taken by one signal to reach to out put of the same gate from input side . Path delay is the time taken by signal to travel from output of one gate to the input of another gate

Delay time expression of CMOS inverter: Average propagation delay

: $$t_p = \frac{(t_{PHL} + t_{PLH})}{2}$$

$$\tau_{PHL} = \frac{C_{load}}{k_n (V_{DD} - V_{T,n})} \left[\frac{2V_{T,n}}{V_{DD} - V_{T,n}} + \ln\left(\frac{4(V_{DD} - V_{T,n})}{V_{DD}} - 1 \right) \right]$$

$$\tau_{PLH} = \frac{C_{load}}{k_p (V_{DD} - |V_{T,p}|)} \left[\frac{2|V_{T,p}|}{V_{DD} - |V_{T,p}|} + \ln\left(\frac{4(V_{DD} - |V_{T,p}|)}{V_{DD}} - 1 \right) \right]$$

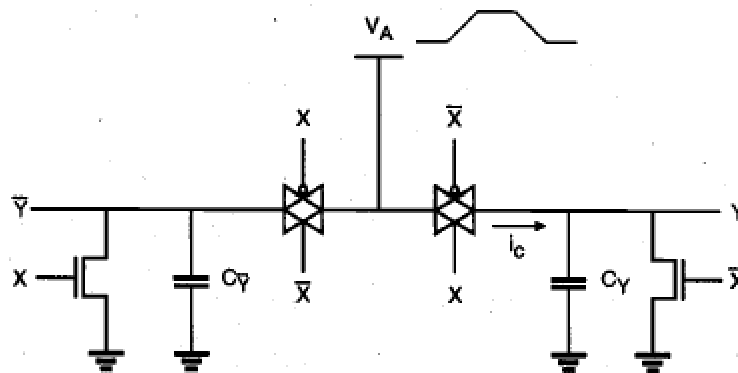
5.3 Adiabatic logic (basic concept)

"Adiabatic" is a term of Greek origin that has spent most of its history associated with classical thermodynamics. It refers to a system in which a transition occurs without energy (usually in the form of heat) being either lost to or gained from the system. In the context of electronic systems, rather than heat, electronic charge is preserved. Thus, an ideal adiabatic circuit would operate without the loss or gain of electronic charge.

The term "adiabatic logic" is used to describe logic families that could theoretically operate without losses.

Here are several important principles that are shared by all of these low-power adiabatic systems. These include only turning switches on when there is no potential difference across them, only turning switches off when no current is flowing through them, and using a power supply that is capable of recovering or recycling energy in the form of electric charge. To achieve this, in general, the power supplies of adiabatic logic circuits have used constant current charging (or an approximation thereto), in contrast to more traditional non-adiabatic systems that have generally used constant voltage charging from a fixed-voltage power supply.

The following circuit is one of the example of adiabatic amplifier . which transfers the complementary input signals to its complementary outputs through CMOS transmission gates



Sample Questions : Module -5

5.1 MCQ

i. If supply voltage decreases by twice then dynamic power across CMOS gate

(a) increases by factor 4 (b) increases by factor 2 (c) decreases by factor 2 (d) decreases by factor 4

ii. If power supply voltage increases gate delay (a) increases (b) remains intact (c) decreases (d) in creases and then saturate

iii . Adiabatic logic is used for (a) delay reduction (b) power reduction (c) No. of transistor reduction (d) All of these

iv. For a long channel device which one is the dominating power
(a) dynamic power (b) short circuit power (c) leakage power (d) subthreshold leakage power

5.2 Short answer type

i. Explain the term dynamic power , short circuit power and leakage power dissipation in a CMOS gate

ii. Define the following Critical path ,arrival time , slack , skew ,set-up time ,hold time , gate delay and path delay

5.3 Long answer type question

i. Find an expression of dynamic power in a CMOS inverter . With an example explain the working of adiabatic logic .