VLSIDESIGN
(20PCC3117)
LectureNotes

B.TECH
(III YEAR–ISEM)
(2023-24)

KALLAM HARANADHAREDDYINSTITUTE OF TECHNOLOGY

IVYearB.Tech.ECE-ISem

VLSI DESIGN

(20PCC3117)

OBJECTIVES

1. Basic characteristics of MOS transistor and examines various possibilities for configuring inverter circuits and aspects of latch-up are considered.
2. Design processes are aided by simple concepts such as stick and symbolic diagrams but the key element is a set of design rules, which are explained clearly.
3. Basic circuit concepts are introduced for MOS processes we can set out approximate circuit parameters which greatly ease the design process..

UNIT I

Introduction and Basic Electrical Properties of MOS Circuits: Introduction to IC technology, Fabrication process: nMOS, pMOS and CMOS.Ids, Aspects of MOS transistor Threshold Voltage, MOS transistor Trans, Output Conductance and Figure of Merit. nMOS Inverter, Pull-up to Pull-down Ratio, MOS inverter driven by another nMOS inverter, and through one or more pass transistors. Alternative forms of pull-up, The CMOS Inverter, Latch-up in CMOS circuits, Bi-CMOS Inverter, Comparison between CMOS and BiCMOS technology.

UNITII

MOS and Bi-CMOS Circuit Design Processes: MOS Layers, Stick Diagrams, Design Rules and Layout, General observations on the Design rules, 2µm Double Metal, Double Poly, MOS/BiCMOS rules, 1.2µmDouble Metal, Double Poly CMOS rules, Layout Diagrams of NAND and NOR gates and CMOS inverter, Symbolic Diagrams-Translation to Mask Form.

UNITIII

Basic Circuit Concepts: Sheet Resistance, Sheet Resistance concept applied to MOS transistors and Inverters, Area Capacitance of Layers, Standard unit of capacitance, some area Capacitance Calculations, The Delay Unit, Inverter Delays, driving large capacitive loads, Propagation Delays, Wiring Capacitances, Choice of layers.

Scaling of MOS Circuits: Scaling models and scaling factors, Scaling factors for device parameters, Limitations of scaling, Limits due to sub threshold currents, Limits on logic levels and supply voltage due to noise and current density. Switch logic, Gate logic..

UNITIV

FPGA Design: FPGA design flow, Basic FPGA architecture, FPGA Technologies, Introduction to FPGA Families.

Design for Testability: Fault types and Models, Controllability and Observability, AdHoc Testable Design Techniques, Scan Based Techniques and Built-In Self Test techniques.

UNIT V

Introduction to Low Power VLSI Design: Introduction to Deep submicron digital IC design**,** Low Power CMOS Logic Circuits: Over view of power consumption, Low–power design through voltage scaling, Estimation and optimization of switching activity, Reduction of switching capacitance.

TEXTBOOKS:

1. Essentials of VLSI Circuits and Systems-Kamran Eshraghian, Douglas and A.Pucknell and Sholeh Eshraghian, Prentice-Hall of India Private Limited, 2005 Edition.
2. CMOS Digital Integrated Circuits Analysis and Design-Sung-Mo Kang, Yusuf Leblebici, Tata Mc Graw-Hill Education, 2003.

REFERENCEBOOKS:

1. Advanced Digital Design with the Verilog HDL, Michael D. Ciletti, Xilinx Design Series, Pearson Education
2. Analysis and Design of Digital Integrated Circuits in Deep submicron Technology, 3rd Edition, David Hodges.

COURSEOUTCOMES

1. Aware the properties of MOS active devices and simple circuits configured when using them and the reason for such encumbrances as ratio rules by which circuits can be inter connected in silicon.
2. Know three sets of design rules with which NMOS and CMOS designs may be fabricated.
3. Aware the scaling factors determining the characteristics and performance of MOS circuits in silicon.
4. Aware the FPGA Architecture, Design flow and different testing techniques to detect the Fault
5. Aware the Low Power CMOS Logic Circuits and know importance of low power consumption.

UNIT-I

ICTechnologies

- Introduction
- MOS
- PMOS
- NMOS
- CMOS

  &
- BiCMOS

  Technologies

BasicElectricalPropertiesof MOS and BiCMOS Circuits

- I_{DS}-V_{DS}relationships
- MOStransistorThreshold Voltage-$V_T$figureof merit-ω0 Transconductance-g_m,

☐ g_{ds};

☐ Passtransistor

☐ NMOS Inverter, Various pull ups, CMOS Inverter analysis and design

☐ Bi-CMOSInverters

INTRODUCTIONTOICTECHNOLOGY

The development of electronics endless with invention of vaccum tubes and associated electronic circuits. This activity termed as vaccum tube electronics, afterward the evolution of solid state devices and consequent development of integrated circuits are responsible for the present status of communication, computing and instrumentation.

- Thefirstvaccumtubediodewasinventedby**johnambrase Fleming**in1904.
- Thevaccumtriodewasinventedby**leedeforest**in1906.

Early developments of the Integrated Circuit (IC) go back to1949.GermanengineerWerner Jacobi filed a patentfor an IC like semiconductor amplifyingdeviceshowingfive transistors on acommonsubstrateina2-stage amplifier arrangement. Jacobi disclosed smallcheap of hearing aids.

Integrated circuits were made possible by experimental discoveries which showed that semiconductor devices could perform the functions of vacuum tubes and by mid-20th-century technology advancements in semiconductor device fabrication.

The integration of large numbers of tiny transistors into a small chip was an enormous improvement over the manual assemblyofcircuits using electronic components.

The integratedcircuits mass productioncapability, reliability, and building-block approachto circuit design ensured the rapid adoption of standardized ICs in place of designs using discrete transistors.

An integrated circuit (IC) is a small semiconductor-based electronic device consisting of fabricated transistors, resistors and capacitors. Integrated circuits are the building blocks of most electronic devices and equipment. An integrated circuit is also known as a chip or microchip.

There are two main advantages of ICs over discrete circuits: cost and performance. Costis low because the chips, with all their components, are printedas a unitby photolithography rather thanbeing constructedonetransistor at atime. Furthermore, much less materialis used to construct a packaged IC die thana discrete circuit. Performance is highsince the components switchquicklyand consume little power (compared to their discrete counterparts) because the components are small and positioned close together. As of 2006, chip areas range from a few square millimeters to around 350 mm^2, with up to 1 million transistors per mm

ICInvention:

Inventor	Year	Circuit	Remark
Fleming	1904 1906	Vacuumtubediode Vacuumtriode	largeexpensive,power-hungry, unreliable
WilliamShockley (Bell labs)	1945	Semiconductorreplacing vacuum tube	--
Bardeen and Brattain and Shockley(Belllabs)	1947	Point Contact transfer resistancedevice“BJT”	Drivingfactorofgrowthof the VLSI technology
WernerJacobi(SiemensAG)	1949	1st ICcontainingamplifying Device 2stage amplifier	Nocommercialuse reported
Shockley	1951	JunctionTransistor	“Practicalformof transistor”
JackKilby (Texas Instruments)	July1958	Integrated Circuits F/F With2-TGermaniumslice and gold wires	FatherofIC design
NoyceFairchild Semiconductor	Dec.1958	IntegratedCircuitsSilicon	“TheMayorofSilicon Valley”
KahngBellLab	1960	FirstMOSFET	Start of new era for semiconductorindustry
Fairchild SemiconductorAnd Texas	1061	FirstCommercial IC	
FrankWanlass (Fairchild Semiconductor)	1963	CMOS	
FedericoFaggin (Fairchild Semiconductor)	1968	Silicon gateICtechnology	Later Joined Intel to lead firstCPUIntel4004in1970 2300T on9mm^2
Zarlink Semiconductors	Recently	M2Acapsulefor endoscopy	takephotographsof digestivetract2/sec.

Moore’sLaw:

- GordonE.Moore-ChairmanEmeritus ofIntelCorporation
- 1965-observedtrendsinindustry-oftransistorsonICsvsreleasedates
- Noticednumberoftransistorsdoublingwithreleaseofeach newICgeneration
- Releasedates(separategenerations)wereall18-24monthsapart

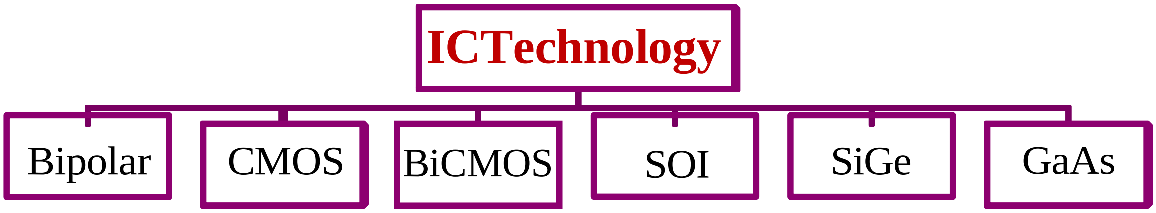
“Thenumberoftransistorsonanintegratedcircuitwilldoubleevery18months”

The levelofintegrationofsilicontechnologyas measured intermsofnumber ofdevicesper IC Semiconductor industry has followed this prediction with surprising accuracy.

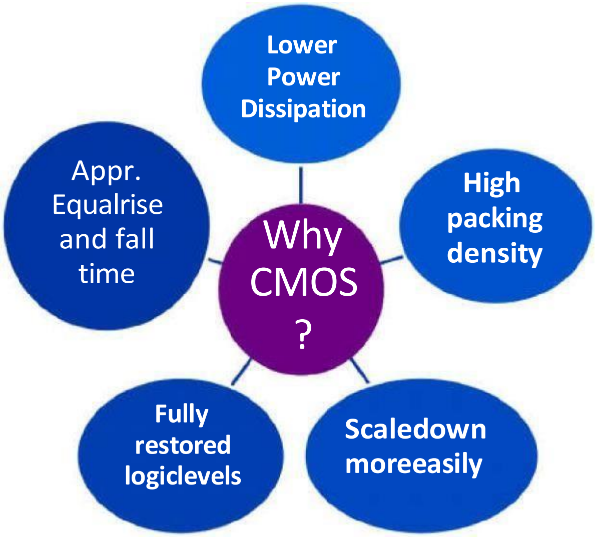
ICTechnology:

- Speed/Powerperformance ofavailabletechnologies
- Themicroelectronicsevolution
- SIARoadmap
- SemiconductorManufacturers2001Ranking

CircuitTechnology



Category	BJT	CMOS
Power Dissipation	Moderate to High	less
Speed	Faster	Fast
Gm	4ms	0.4ms
Switch implementation	poor	Good
Techn ology improvement	slower	Faster



ScaleofIntegration:

- **Smallscaleintegration(SSI)--1960**

  Thetechnologywasdevelopedbyintegratingthenumber oftransistorsof1-100 on a single chip. Ex: Gates, flip-flops, op-amps.

- **Mediumscaleintegration(MSI)--1967**

  Thetechnologywasdeveloped byintegrating thenumber oftransistorsof100- 1000 on a single chip. Ex: Counters, MUX, adders, 4-bit microprocessors.

- **Largescaleintegration(LSI)--1972**

  Thetechnologywasdeveloped byintegratingthenumber oftransistorsof1000- 10000 on a single chip. Ex:8-bitmicroprocessors,ROM,RAM.

- **Verylargescaleintegration(VLSI)-1978**

  Thetechnologywasdeveloped byintegrating thenumber oftransistorsof10000- 1Million on a single chip. Ex:16-32 bit microprocessors, peripherals, complimentary high MOS.

- **Ultralargescaleintegration(ULSI)**

  Thetechnologywasdeveloped byintegratingthenumber oftransistorsof1Million- 10 Millions on a single chip. Ex: special purpose processors.

- **Giantscaleintegration(GSI)**

  Thetechnologywasdeveloped byintegrating thenumber oftransistorsofabove10 Millions on a single chip. Ex: Embedded system, systemon chip.

- ✓ Fabricationtechnologyhasadvancedtothepointthat wecanput acompletesystemona single chip.
- ✓ Singlechip computercanincludeaCPU,bus,I/Odevicesandmemory.
- ✓ This reduces the manufacturing cost than the equivalent boardlevel systemwithhigher performance and lower power.

MOSTECHNOLOGY:

MOS technologyisconsidered asone ofthe very important and promising technologies in the VLSI design process. The circuit designs are realized based on pMOS, nMOS, CMOS and BiCMOS devices.

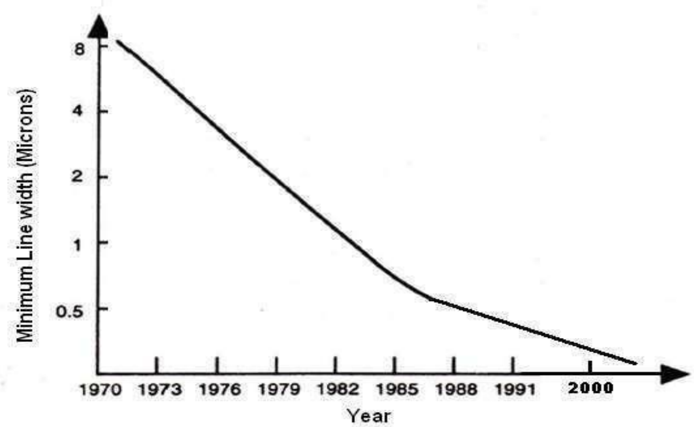
The pMOS devices are based on the p-channel MOS transistors.Specifically,thepMOSchannel is part of a n-type substrate lying between twoheavily doped p+ wells beneath the source and drain electrodes. Generally speaking, a pMOS transistor is only constructedin consort with an NMOS transistor.

The nMOS technology and design processes provide an excellent background for other technologies. In particular, some familiarity with nMOS allows a relatively easy transition to CMOS technology and design.

The techniques employed in nMOS technologyfor logic designare similar to GaAs technology.. Therefore, understanding the basics ofnMOS design will help inthe layout ofGaAs circuits

In addition to VLSI technology, the VLSI design processes also provides a new degree of freedom for designers which helps for the significant developments. With the rapid advances in technologythe the size ofthe ICs is shrinking and the integration density is increasing.

Theminimumlinewidthofcommercialproductsovertheyearsisshowninthegraphbelow.



Thegraphshowsasignificant decrease inthesize ofthechip inrecent yearswhichimplicitly indicates the advancements in the VLSI technology.

MOSTransistorSymbol:

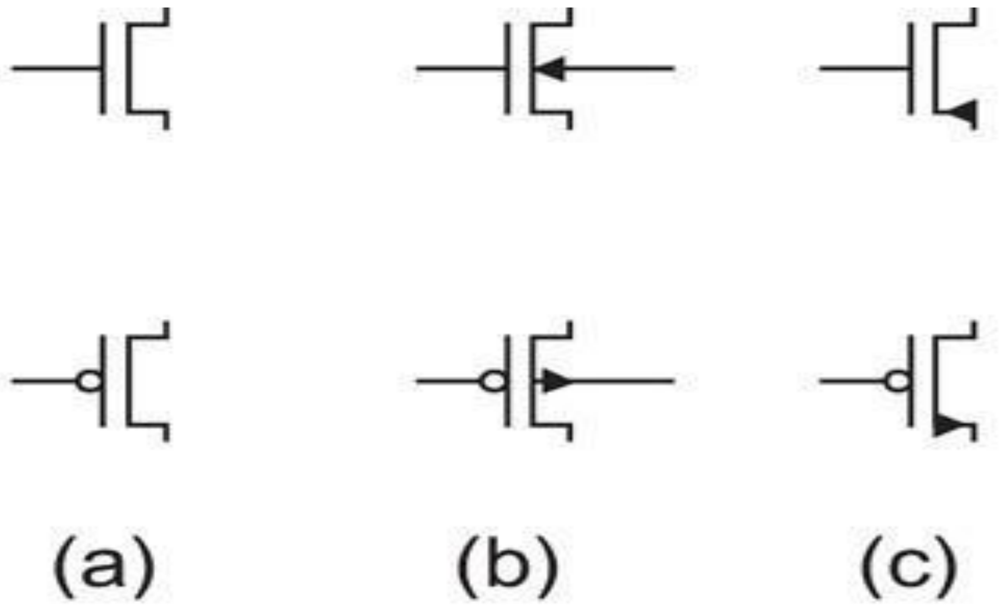


FIG 2.1 MOS transistor symbols

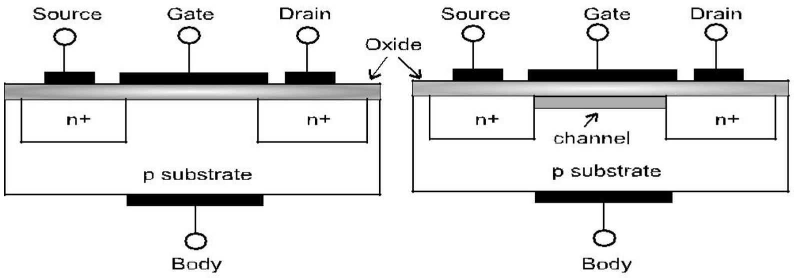
ENHANCEMENTANDDEPLETIONMODEMOSTRANSISTORS

MOS Transistors are built on a silicon substrate. Silicon which is a group IV material is the eighth most commonelement inthe universe bymass, but veryrarelyoccurs asthe pure free element in nature. It is most widely distributed in dusts, sands, planetoids, and planets as various forms of silicon dioxide (silica) or silicates. It forms crystal lattice with bonds to four neighbours. Silicon is a semiconductor. Pure silicon has no free carriers and conducts poorly. But adding dopants to silicon increases its conductivity. If a group V material i.e. an extra electron is added, it forms an n-type semiconductor. If a group III material i.e. missing electron pattern is formed (hole), the resulting semiconductor is called a p-type semiconductor.

A junction between p-type and n-type semiconductor forms a conduction path. Source and Drain ofthe MetalOxide Semiconductor (MOS) Transistor is formed bythe “doped” regions onthe surface of chip. Oxide layer is formed by means of deposition of the silicon dioxide (SiO_2) layer which forms as an insulator and is a verythin pattern. Gateof the MOS transistor is the thin layer of “polysilicon (poly)”; used to apply electricfield to the surface of silicon between Drain and Source, to forma “channel” of electrons or holes. Controlby the Gate voltage is achieved by modulating the conductivityofthe semiconductor region just below the gate. This region is knownas thechannel.

The Metal–Oxide–Semiconductor Field Effect Transistor (MOSFET) is a transistor which is a voltage-controlled current device,in which current at two electrodes, drain and source is controlled by the action of an electric field at another electrode gate havingin-between semiconductor and a verythin metaloxide layer. It is used for amplifying or switching electronicsignals.

The Enhancement and Depletion mode MOS transistors are further classified as N-type named NMOS (or N-channel MOS) and P-type named PMOS (or P-channel MOS) devices.Figure 1.5 shows the MOSFETs along with their enhancement and depletion modes.



Fi

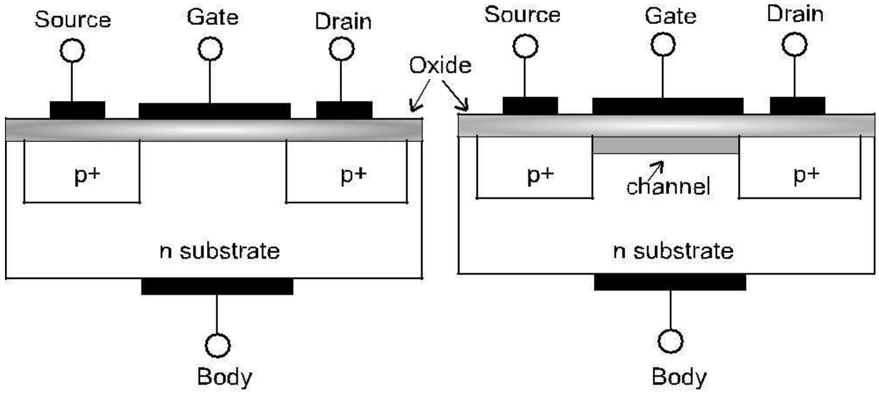


Figure1.5:(c)EnhancementP-typeMOSFET(d) DepletionP-typeMOSFET

The depletion mode devices are doped so that a channel exists even with zero voltage from gate to source during manufacturing of the device. Hence the channel always appears in the device. To controlthe channel, a negative voltage is applied tothe gate(for anN-channeldevice), depleting the channel, which reduces the current flow throughthe device. In essence, the depletion-mode device is equivalent to a closed (ON) switch, while the enhancement-mode device does not have the built in channel and is equivalent to an open (OFF) switch. Due to the difficulty of turning off the depletion mode devices, they are rarely used

Working ofEnhancementModeTransistor

The enhancement mode devicesdo not have the in-built channel. Byapplying the required potentials, the channelcan be formed. Also for the MOS devices, there is a threshold voltage (V_t), below which not enough charges will be attracted for the channel to be formed. This threshold voltage for a MOS transistor is a function ofdoping levels and thickness ofthe oxide layer.

Case1:V_{gs}=0VandV_{gs}<V_t

The device is non-conducting, whenno gatevoltage is applied (V_{gs}= 0V)or(V_{gs}< V_t) and also drain to source potential V_{ds} = 0. With an insufficient voltage on the gate to establish the channelregion as N-type, there will be no conduction between the source and drain. Since there is no conducting channel, there is no current drawn, i.e. I_{ds}= 0, and the device is said to be inthe **cut-off region**. This is shown in the Figure 1.7 (a).

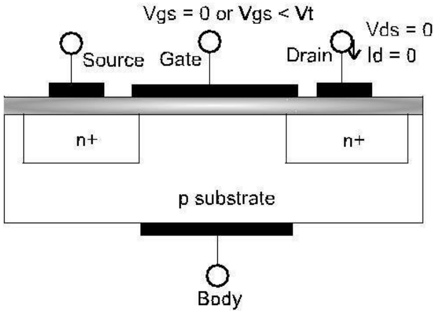


Figure 1.7:(a)Cut-offRegion

Case 2:Vgs>Vt

When a minixmum voltage greater than the threshold voltage V_t (i.e. V_{gs} > V_t) is applied, a high concentration of negative charge carriers forms an inversion layer located by a thin layer next to the interface between the semiconductor and the oxide insulator. Thisforms a channel between the source and drain ofthe transistor. This is shown in the Figure 1.7 (b).

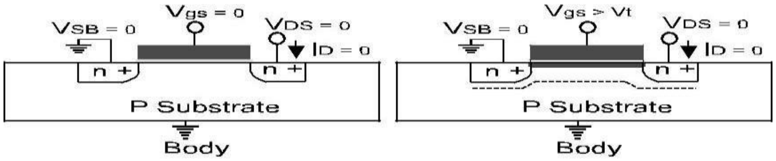


Figure1.7:(b)FormationofaChannel

A positive V_{ds} reverse biases the drain substrate junction, hence the depletion region around the drain widens, and since the drain is adjacent to the gate edge, the depletion region widens in the channel. This is shown in Figure 1.7 (c). This results in flow of electron from source todrain resulting in current Ids.. The device is said to operate in **linear region** during this phase. Further increase in $V_{ds,}$ increases the reverse bias on the drain substrate junction in contact with the inversion layer which causes inversion layer density to decrease. This is shown in Figure 1.7 (d). The point at which the inversion layer density becomes very small (nearly zero) at the drain end is termed pinch- off. The value of V_{ds} at pinch-off is denoted as $V_{ds,sat}$. This is termed as **saturation region** for the MOS device. Diffusion current completes the path from source to drain in this case, causing the channelto exhibit a high resistance and behaves as a constant current source.

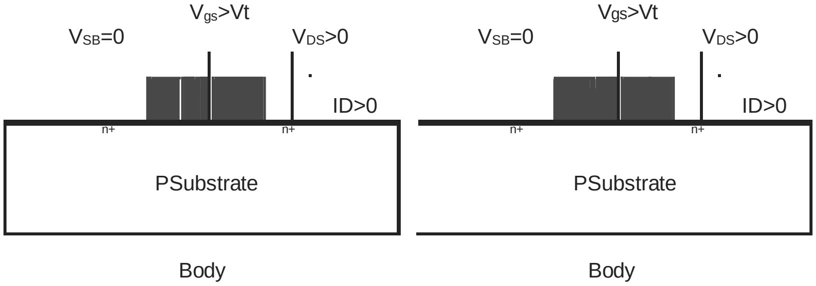


Figure1.7:(c)LinearRegion.(d)SaturationRegion

TheMOSFET$I_D$versusV_{DS}characteristics(V-I Characteristics)isshownintheFigure1.8.ForV_{GS} $< V_t$, $I_D = 0$ and device is in cut-off region. As V_{DS} increases at a fixed V_{GS}, I_D increases in the linear region due to the increased lateral field, but at a decreasing rate since the inversion layer density is decreasing. Once pinch-off is reached, further increase in V_{DS} results in increase in I_D; due to the formation of the high field region which is very small. The device starts in linear region, and moves into saturation region at higher V_{DS}.

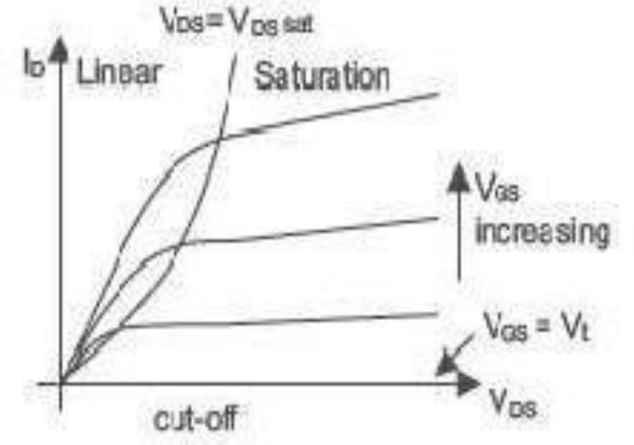


Figure 1.8: MOS V-I Characteristics

NMOS FABRICATION

Thefollowingdescriptionexplainsthebasicstepsusedintheprocessoffabrication.

(a) The fabricationprocessstartswiththeoxidationofthesiliconsubstrate. It is shown in the Figure 1.9 (a).

(b) Arelativelythicksilicondioxide layer, also calledfieldoxide, iscreatedonthesurfaceofthe substrate. This is shown in the Figure 1.9 (b).

(c) Then,thefieldoxide isselectivelyetchedtoexposethesiliconsurfaceonwhichtheMOS transistor will be created. This is indicated in the Figure 1.9 (c).

(d) Thisis followed bycoveringthesurfaceofsubstratewithathin, high-qualityoxide layer, which will eventually form the gate oxide of the MOStransistorasillustratedinFigure1.9(d).

(e) Ontopofthethinoxide, alayer ofpolysilicon(polycrystalline silicon) is deposited as is shown in the Figure 1.9 (e). Polysilicon is used both as gate electrode material for MOS transistors and also as an interconnect medium in silicon integrated circuits. Undoped polysilicon has relatively high resistivity. The resistivityofpolysilicon can be reduced, however, bydoping it with impurityatoms.

(f) After deposition, the polysilicon layer is patterned and etched to form the interconnects and the MOS transistor gates. This is shown in Figure 1.9 (f).

(g) The thin gate oxide not covered by polysilicon is also etched along, which exposes the bare silicon surface on which the source and drain junctions are to be formed (Figure 1.9 (g)).

(h) The entire silicon surface is then doped with high concentration of impurities, either through diffusion or ion implantation (in this case with donor atoms to produce n-type doping). Diffusion is achievedbyheatingthewafer to ahightemperatureandpassingthegascontainingdesired impurities over the surface. Figure 1.9 (h) shows that the doping penetrates the exposed areas on the silicon surface, ultimately creating two n-type regions (source and drain junctions) in the p-type substrate. The impuritydoping also penetratesthe polysiliconon the surface, reducing itsresistivity.

(i) Once the source and drain regions are completed, the entire surface is again covered with an insulating layer of silicon dioxide, as shown in Figure 1.9 (i).(*j*) The insulating oxide layer is then patterned in orderto provide contact windows for the drain and source junctions, as illustrated in Figure 1.9 (j).

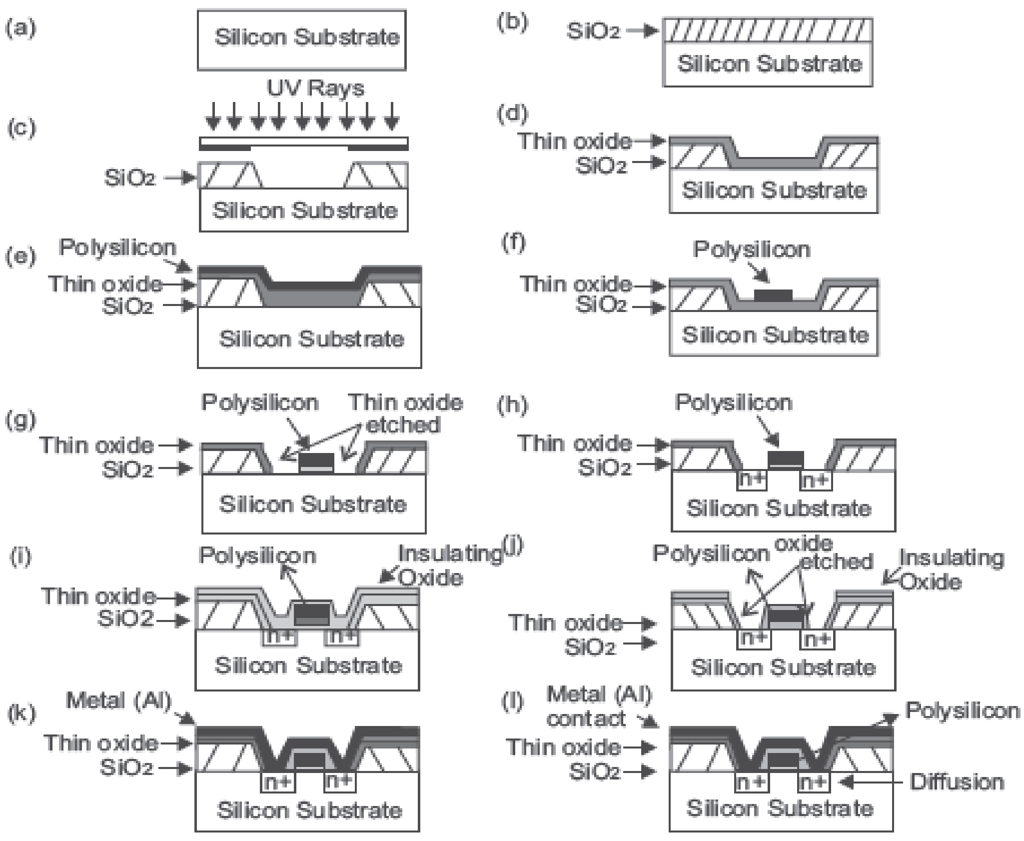


Figure 1.9: Fabrication Process of NMOS Device

CMOSFABRICATION:

CMOSfabricationcanbeaccomplished using eitherofthethreetechnologies:

- N-welltechnologies/P-welltechnologies
- Twinwelltechnology
- SiliconOnInsulator(SOI)

The fabrication of CMOS can be done by following the below shown twentysteps, bywhich CMOS can be obtained by integrating boththe NMOS and PMOStransistorsonthe same chip substrate. For integrating these NMOS and PMOS devices onthe same chip, special regions called as wells or tubs are required in which semiconductor type and substratetype are opposite to each other.

A P-well has to be created on a N-substrate or N-well has to be created on a P-substrate. In this article, the fabrication of CMOS is described using the P-substrate, in which the NMOS transistor is fabricated on a P-type substrate and the PMOS transistor is fabricated in N-well.

Thefabricationprocessinvolvestwentysteps,whichareasfollows:

N-WellProcess

Step1:Substrate

Primarily,starttheprocesswithaP-substrate.

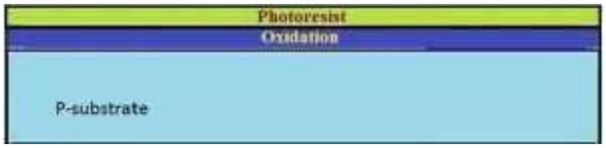
Step2: Oxidation

Theoxidationprocess isdonebyusing high-purityoxygenand hydrogen, whichareexposed inan oxidation furnace approximately at 1000 degree centigrade.

Step3: Photoresist

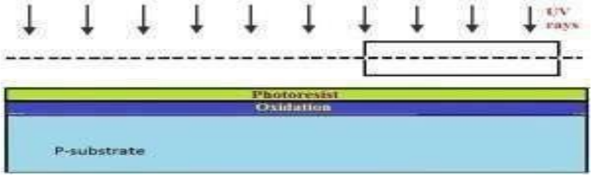
Alight-sensitivepolymerthatsoftenswheneverexposedtolightiscalledasPhotoresistlayer.

Itis formed.



Step4:Masking

ThephotoresistisexposedtoUVraysthroughtheN-wellmask

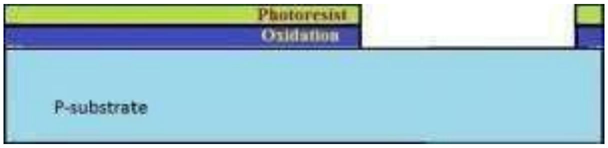


Step5:Photoresistremoval

Apartofthephotoresistlayerisremovedbytreatingthewaferwiththebasicoracidicsolution.

Step6:RemovalofSiO2usingacidetching

TheSiO2oxidationlayer isremovedthroughtheopenarea madebytheremovalofphotoresist using hydrofluoric acid.



Step7:Removalofphotoresist

Theentirephotoresistlayerisstrippedoff, asshowninthebelowfigure.

Step8: Formation oftheN-well

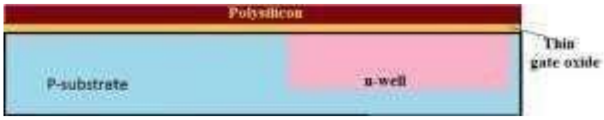
ByusingionimplantationordiffusionprocessN-wellisformed.

Step9:RemovalofSiO2

Usingthehydrofluoricacid, theremainingSiO2 isremoved.

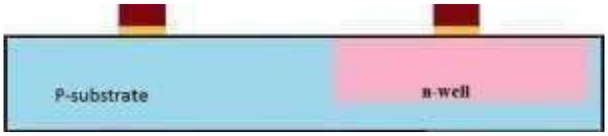
Step10:Depositionofpolysilicon

ChemicalVaporDeposition(CVD)processisusedtodepositaverythinlayerofgateoxide.



Step11:Removingthelayerbarringa smallarea forthe Gates

Exceptthetwo smallregionsrequired forformingtheGatesofNMOSandPMOS,theremaining layer is stripped off.



Step12:Oxidation process

Next,anoxidationlayer is formedonthis layerwithtwo smallregions fortheformationofthegate terminals of NMOS and PMOS.

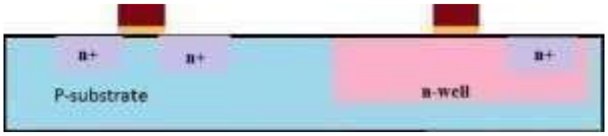
Step13:Masking andN-diffusion

ByusingthemaskingprocesssmallgapsaremadeforthepurposeofN-diffusion.

Then-type(n+)dopantsarediffusedorionimplanted,andthethreen+areformed fortheformation of the terminals of NMOS.

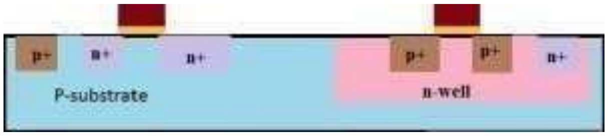
Step14:Oxidestripping

Theremainingoxidationlayerisstrippedoff.



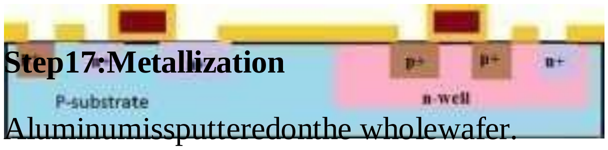
Step15:P-diffusion

Similar totheaboveN-diffusionprocess,theP-diffusionregionsarediffusedto formtheterminalsof the PMOS.



Step16:Thickfield oxide

Athick-fieldoxideisformedinallregionsexcepttheterminalsofthePMOSandNMOS.



Step17:Metallization

Aluminumissputteredonthe wholewafer.

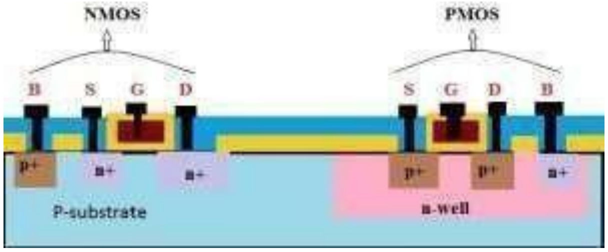
Step18: Removalofexcessmetal

Theexcessmetalisremovedfromthewaferlayer.

Step19: Terminals

TheterminalsofthePMOSandNMOS aremadefromrespectivegaps.

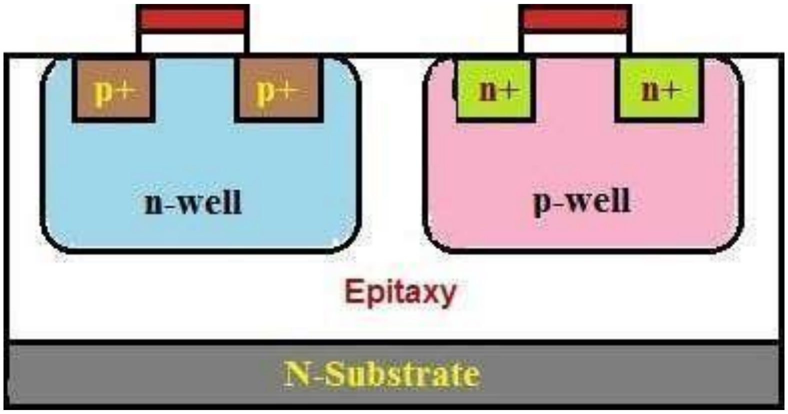
Step20:AssigningthenamesoftheterminalsoftheNMOSand PMOS



Fabrication ofCMOSusingP-wellprocess

Among all the fabrication processes of the CMOS, N-well process is mostly used for the fabrication of the CMOS. P-well process is almost similar to the N-well. But the only differencein p-well process is that it consists of a main N-substrate and, thus, P-wells itself acts as substrate for the N- devices.

Twintub-CMOSFabricationProcess



In this process, separate optimization of the n-type and p-type transistorswill be provided. The independent optimization of Vt, body effect and gain of the P-devices, N-devices can be made possible with this process.

DifferentstepsofthefabricationoftheCMOS usingthetwintubprocessare asfollows:

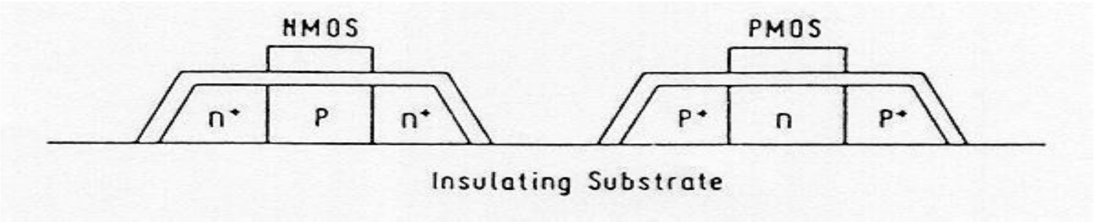
- Lightlydoped n+orp+substrateistakenand,toprotectthelatchup,epitaxiallayerisused.
- Thehigh-puritycontrolledthicknessofthe layersofsiliconaregrownwithexactdopant concentrations.
- Thedopantanditsconcentration inSiliconareusedtodetermineelectricalproperties.
- Formationofthetub
- Thinoxideconstruction

- Implantationofthe sourceanddrain
- Cutsfor makingcontacts
- Metallization

Byusingthe above stepswe can fabricateCMOSusing twintubprocessmethod.

Silicon-on-Insulator(SOI)CMOSProcess

Rather than using silicon as the substrate material, technologists have sought to use an insulating substrateto improve process characteristics such as speed and latch-up susceptibility. The SOI CMOS technology allows the creation of independent, completely isolated nMOS and pMOS transistors virtually side-by-side on an insulating substrate. The main advantages of this technology arethehigher integrationdensity(becauseoftheabsenceofwellregions), completeavoidanceofthe latch-upproblem,and lowerparasiticcapacitancescomparedtotheconventionalp&n-wellortwin- tub CMOS processes. A cross-section of nMOS and pMOS devicesusing SOI processis shownbelow.

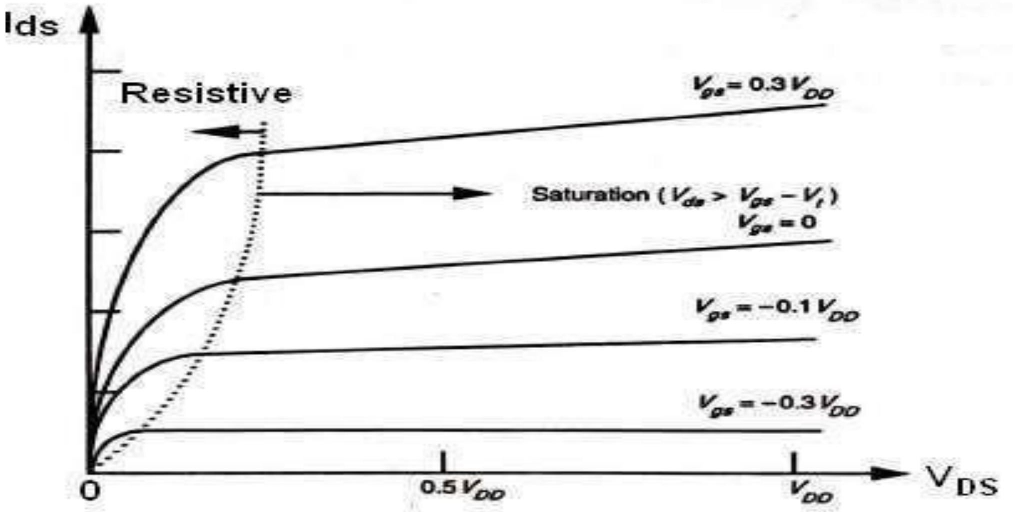


TheSOI CMOSprocess isconsiderablymorecostlythanthestandardp&n-wellCMOSprocess. Yet theimprovementsofdeviceperformance and theabsenceoflatch-up problemscan justify its use, especially for deep-sub-micron devices.

BasicElectricalPropertiesofMOSandBiCMOScircuits

I_D-V_{DS}CharacteristicsofMOSTransistor :

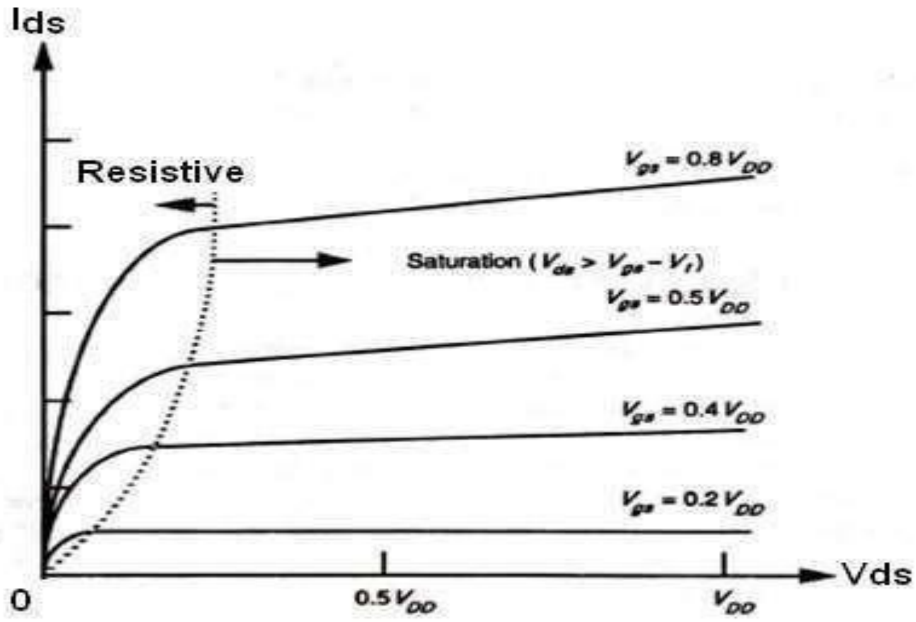
The graph below shows the ID Vs VDS characteristics of an n- MOS transistor for several values of VGS .It is clear that there are two conduction states when the device is ON. The saturated stateand the non-saturated state. The saturated curve is the flat portion and defines the saturation region. For Vgs < VDS + Vth, the nMOS device is conducting and ID is independent ofVDS. For Vgs > VDS+ Vth, thetransistor is inthe non-saturationregionand the curve is a halfparabola. Whenthetransistor is OFF (Vgs < Vth), then ID is zero for anyVDS value.



(a) Depletion mode device

Theboundaryofthesaturation/non-saturation bias statesisapoint seen for eachcurve inthegraphas the intersection of the straight line of the saturated region with the quadratic curve of the non- saturated region. This intersection point occurs at the channel pinch off voltage called VDSAT. The diamond symbol marks the pinch-off voltage VDSAT for each value of VGS. VDSAT is defined as the minimumdrain-source voltagethat isrequired to keepthetransistorinsaturation for a givenVGS .In the non-saturated state, the drain current initially increases almost linearly from the origin before bending ina parabolic response. Thus the name ohmic or linear forthe non- saturatedregion.

Thedraincurrentinsaturation isvirtuallyindependentofVDSandthetransistoractsasacurrent source. This is because there is no carrier inversion at the drain region of the channel. Carriers are pulled into the high electric field of the drain/substrate pn junction and ejected out of the drain terminal.



(b). Enhance mode device

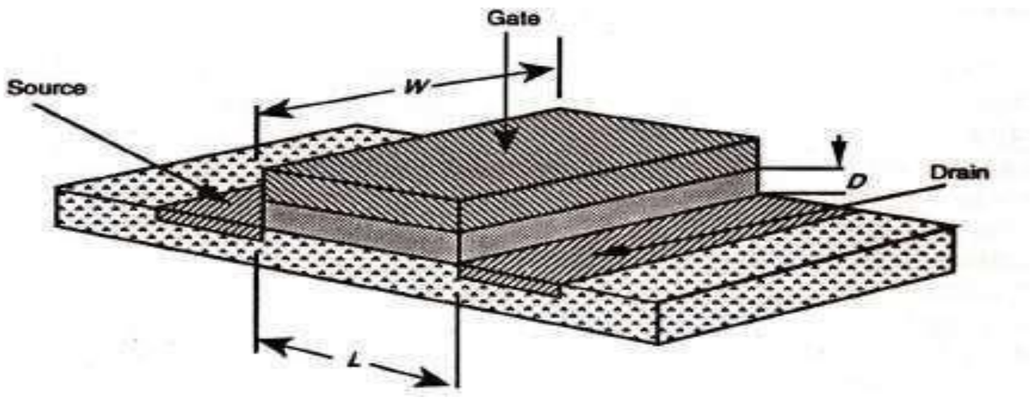
Drain-to-SourceCurrentI_{DS}VersusVoltageV_{DS}Relationships:

The working ofa MOStransistor is based onthe principle thatthe useofa voltageonthe gateinduce a charge in the channel between source and drain, which may then be caused to move from sourceto drain under the influence of an electric field created by voltage Vds appliedbetween drain and source. Sincethecharge induced isdependent onthegateto sourcevoltageVgsthenIds isdependent on both Vgs and Vds.

Letus considerthe diagram belowin which electrons will flow source to drain .So,the drain current is given by

Charge induced in channel (Qc) Ids =-Isd = Electron transit time(τ) Length ofthe channel Where the transit time is given by τsd =-----------------------------

Velocity(v)



Butvelocity$v=\mu E_{ds}$

Whereμ=electronorholemobilityand E_{ds}=Electricfieldalso,$E_{ds}=V_{ds}/L$

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

Thetypicalvaluesofμ atroomtemperaturearegivenbelow.

$$\mu_n \doteq 650\ \text{cm}^2/\text{V sec (surface)}$$
$$\mu_p \doteq 240\ \text{cm}^2/\text{V sec (surface)}$$

Non-saturatedRegion:

Let us consider the I_d vs V_d relationships in the non-saturated region .The charge induced in the channel due to due to the voltage difference between the gate and the channel, Vgs (assuming substrateconnected to source). Thevoltagealong thechannel varies linearlywithdistance *X* fromthe source due to the IR drop in the channel .In the non-saturated state the average value is Vds/2. Also the effective gate voltage Vg = Vgs – Vt where Vt, is the threshold voltage needed to invert the charge under the gate and establish the channel.

Hencetheinducedchargeis$Q_c=E_g\varepsilon_{ins}\varepsilon_o W.L$ Where

E_g=averageelectricfieldgateto channel

ε_{ins}=relativepermittivityofinsulationbetweengateand channelε_o=permittivity

$$E_g = \frac{\left((V_{gs} - V_t) - \frac{V_{ds}}{2}\right)}{D}$$

HereDisthethicknessoftheoxidelayer.Thus

$$Q_c = \frac{WL\varepsilon_{ins}\varepsilon_0}{D}\left((V_{gs} - V_t) - \frac{V_{ds}}{2}\right)$$

So,bycombining theabovetwoequations,weget

$$I_{ds} = \frac{\varepsilon_{ins}\varepsilon_0\mu}{D}\frac{W}{L}\left((V_{gs} - V_t) - \frac{V_{ds}}{2}\right)V_{ds}$$

ortheaboveequationcanbewrittenas

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

Inthe non-saturatedorresistiveregionwhereVds<Vgs–Vtand

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

Generally,aconstantβisdefinedas

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

Sothat,theexpressionfordrain–sourcecurrentwillbecome

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

Thegate/channelcapacitanceis

$$C_g = \frac{\varepsilon_{ins}\varepsilon_0 WL}{D}\ \text{(parallel plate)}$$

Hencewecanwriteanotheralternativeformforthedraincurrentas

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

Sometimeitisalsoconvenienttousegate–capacitanceperunitarea,C_gSo,thedraincurrentis

$$I_{ds} = C_0\mu\frac{W}{L}\left((V_{gs} - V_t)V_{ds} - \frac{V_{ds}^2}{2}\right)$$

This istherelationbetweendraincurrentanddrain-sourcevoltageinnon-saturatedregion.

SaturatedRegion

Saturation begins when Vds = Vgs - V, since at this point the IR drop in the channel equals the effective gate to channel voltage at the drain and we may assume that the current remains fairly constant as Vds increases further. Thus

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

orwecanalsowritethat

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

oritcanalsobewrittenas

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

or

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

The expressions derived above for I_{ds} hold for both enhancement and depletion mode devices. Here thethreshold voltage for the nMOS depletion mode device (denoted as V_{td}) is negative.

MOSTransistorThreshold VoltageVt:

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

ThethresholdvoltageV$_t$maybeexpressedas:

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

where Q$_D$=thechargeperunit area inthedepletionlayerbelowtheoxideQss=chargedensityat Si: SiO$_2$ interface

C$_o$ =Capacitanceperunitarea.

Φns=workfunctiondifferencebetweengateandSi

Φ$_{fN}$=FermilevelpotentialbetweeninvertedsurfaceandbulkSi

For polynomial gate and silicon substrate, the value of Φ$_{ns}$ is negative but negligible and the magnitude and sign of Vt are thus determined by balancing the other terms in the equation. To evaluate the Vt the other terms are determined as below.

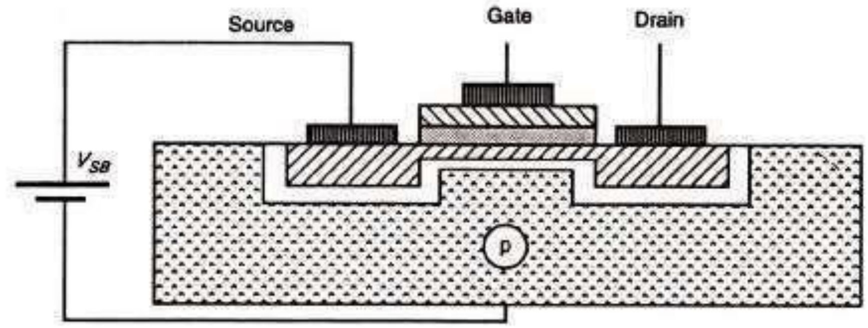
$$Q_B = \sqrt{2\varepsilon_0 \varepsilon_{Si} q N (2\phi_{fN} + V_{SB})} \text{ coulomb/m}^2$$

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

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

BodyEffect:

Generallywhile studying the MOStransistors it is treatedas athreeterminaldevice. But, the bodyof the transistor is also an implicit terminal which helps to understand the characteristics of the transistor. Considering the bodyofthe MOS transistor as a terminal is known as the bodyeffect. The potentialdifference between the source and the body(Vsb) affects the threshold voltage of the transistor. In many situations, this Body Effect is relatively insignificant, so we can (unless **otherwise** stated) ignore the Body Effect. But it is not always insignificant, in some cases it can have a tremendous impact on MOSFET circuit performance.



Bodyeffect-nMOS device

Increasing Vsb causes the channel to be depleted of charge carriers and thus the threshold voltage is raised. Change in Vt is given by $\Delta Vt = \gamma.(Vsb)^{1/2}$ where γ is a constant which depends on substrate doping so that the more lightlydoped the substrate, the smaller will be the bodyeffect

Thethresholdvoltagecanbewrittenas

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

WhereVt(0)isthethresholdvoltageforV_{sd}=0

For n-MOS depletion mode transistors ,the body voltage values at different V_{DD} voltages are given below.

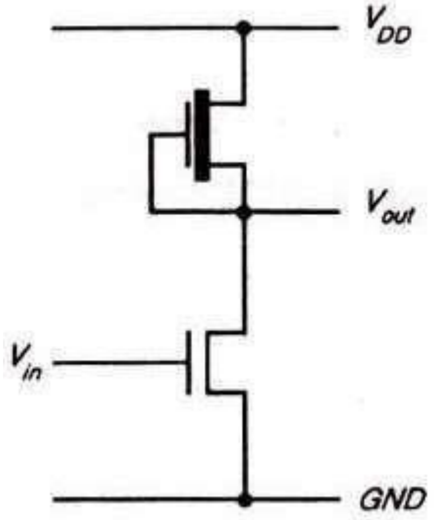
V_{SB} = 0 V ; Vsd = -0.7V_{DD} (= - 3.5 V for V_{DD} =+5V) V_{SB} = 5 V ; Vsd = -0.6V_{DD} (= - 3.0 V for V_{DD} =+5V)

nMOSINVERTER:

An invertercircuit isa veryimportant circuit for producing a complete range oflogic circuits. This is needed for restoring logic levels, for Nand and Nor gates, and for sequential and memory circuits of various forms .

A simple inverter circuit can be constructed using a transistor with source connected to ground and a load resistor of connectedfrom the drain to the positivesupply rail V_{DD}· The outputis taken from the drain and the input applied between gate and ground .

But, during the fabrication resistors are not conveniently produced on the silicon substrate and even small values of resistors occupyexcessively large areas .Hence some other formof load resistance is used. A more convenient wayto solve this problem is to use a depletion mode transistor as the load, as shown in Fig. below.



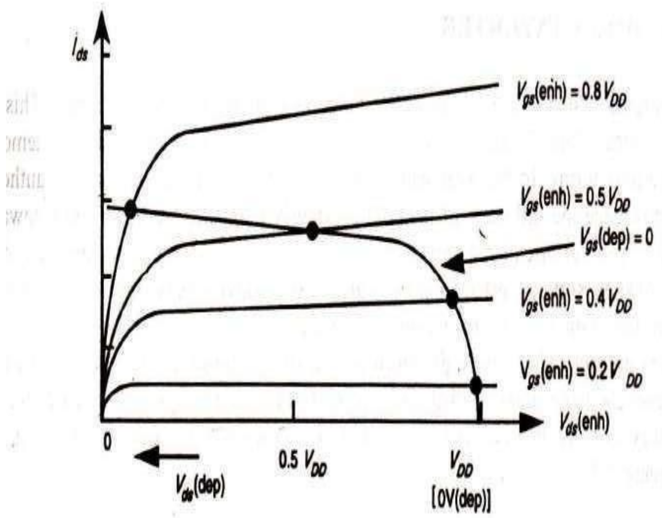
Thesalientfeaturesofthen-MOSinverter are

- Forthedepletionmodetransistor,thegateisconnectedtothesourcesoitisalwayson.
- Inthisconfigurationthedepletionmodedevice iscalledthepull-up(P.U)andtheenhancement mode device the pull-down (P.D) transistor.
- Withnocurrentdrawnfromtheoutput,thecurrentsIdsforbothtransistorsmustbeequal.

nMOSInvertertransfercharacteristic.

Thetransfer characteristic isdrawnbytakingVds onx-axisandIdsonY-axis forbothenhancement and depletion mode transistors. So,to obtain the inverter transfer characteristic for

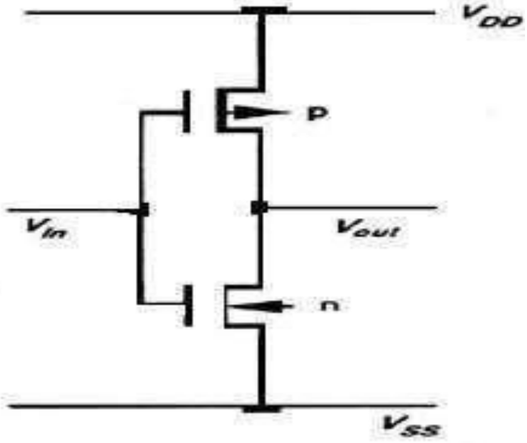
Vgs = 0 depletion mode characteristic curve is superimposed on the family of curves for the enhancement mode device and from the graph it can be seen that , maximum voltage across the enhancement mode device corresponds to minimum voltage acrossthe depletion mode transistor.



From the graph it is clear that as Vin(=Vgs p.d. transistor) exceeds the Pulldown threshold voltage current begins to flow. The output voltage Vout thus decreases and the subsequent increases in Vin willcause the Pulldown transistor to come out ofsaturation and become resistive.

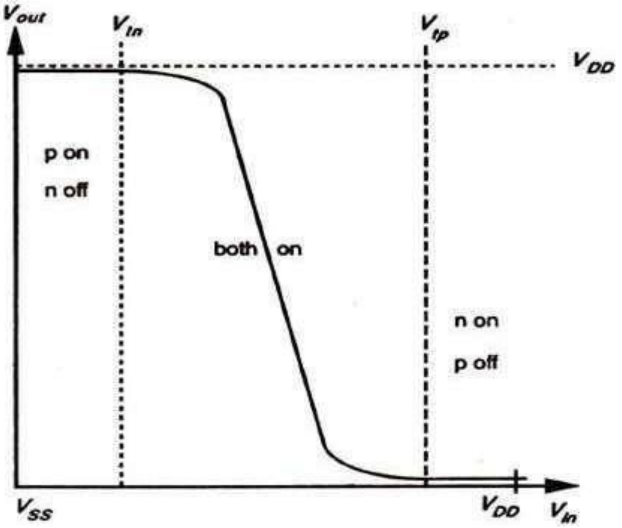
CMOS Inverter:

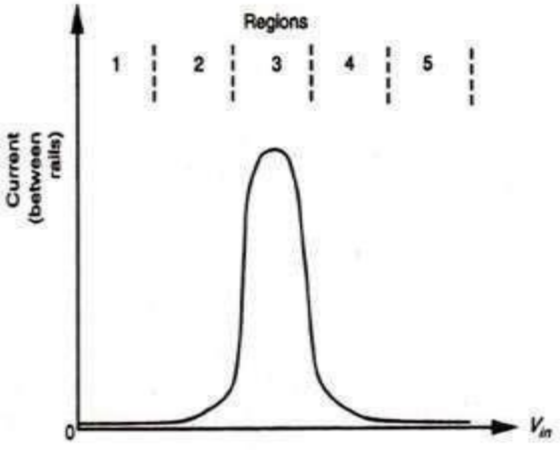
The inverter is the very important part of all digital designs. Once its operation and properties are clearly understood, Complex structures like NAND gates, adders, multipliers, and microprocessors can also be easily done. The electrical behavior of these complex circuits can be almost completely derived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



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

Considering the static conditions first, it may be Seen that in region 1 for which Vi,. = logic 0, we have the p-transistor fully turned on while the n-transistor is fully turned off. Thus no current flows throughthe inverter and the output is directlyconnected to VDD throughthe p-transistor.





Hence the output voltage is logic 1 . In region 5 , Vin = logic 1 and the n-transistor is fullyon while the p-transistor is fullyoff. So, no current flows and logic 0 appears at the output.

In region 2 the inputvoltage has increased to a level which justexceeds the threshold voltage of the n-transistor. The n-transistor conducts and has a large voltage between source and drain; so it is in saturation. The p-transistor is also conducting butwith only a small voltage across it, it operates in theunsaturatedresistiveregion. Asmallcurrent nowflowsthroughthe inverter from VDDto VSS.If we wish to analyze the behavior in this region, we equate the p-device resistive region currentwith the n-device saturation current and thus obtain the voltage and current relationships.

Region 4 is similar to region 2 but with the roles of the p- and n-transistors reversed.However, the current magnitudes in regions 2 and 4 are small and most of the energyconsumed in switching from one stateto the other is due to the larger current which flows inregion3.

Region 3 is the region in which the inverter exhibits gain and in which both transistors are in saturation.

Thecurrentsineachdevicemustbethesame,sincethetransistorsareinseries.So,wecanwritethat

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

where

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

and

$$I_{dsn} = \frac{\beta_n}{2}(V_{in} - V_{tn})^2$$

Since bothtransistorsareinsaturation, theyact ascurrent sourcessothattheequivalent circuit in this region is two current sources in series between VDD and Vss with the output voltage coming from their common point. The region is inherently unstable in consequence and the changeover from one logic level to the other is rapid.

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

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

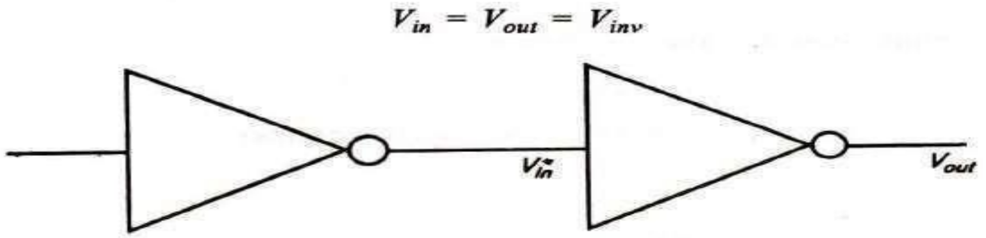


Fig.(a).Inverter drivenbyanotherinverter.

Forequalmarginsaroundthe inverterthreshold, weset Vinv=0.5V_{DD}·Atthispoint both transistors are in saturation and we can write that

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

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

and in the enhancement mode

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

Equating (since currents are the same) we have

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

where $W_{p.d}$, $L_{p.d}$,$W_{p.u.}$ and$L_{p.u}$arethewidths and lengthsofthepull-downandpull-up transistors respectively.

So,wecanwritethat

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

we have

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

whence

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

Thetypical,valuesforVt,VinvandVtdare

$$V_t = 0.2V_{DD};\ V_{td} = -\ 0.6V_{DD}$$
$$V_{inv} = 0.5V_{DD}\ \text{(for equal margins)}$$

Substitutingthesevalues intheaboveequation,weget

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

Here

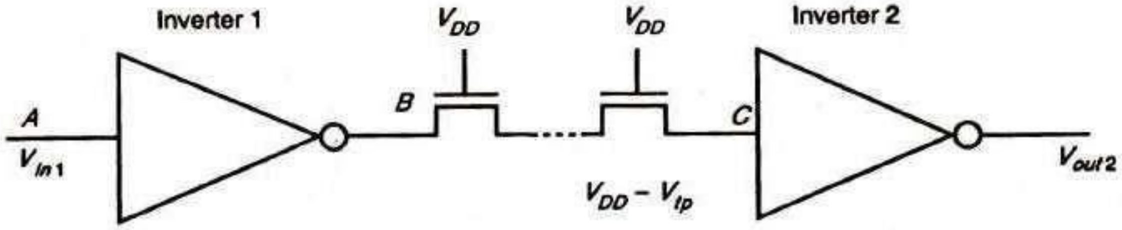
$$\sqrt{Z_{p.u.}/Z_{p.d.}} = 2$$

So,weget

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

Thisistheratio forpull-upto pulldownratio foraninverterdirectlydrivenbyanother inverter. **Pull -Up to Pull-Down ratio for an nMOS Inverter driven through one or more Pass Transistors**

Letusconsideranarrangementinwhichtheinputtoinverter 2comesfromtheoutputofinverter1 but passesthroughoneormorenMOStransistorsasshowninFig. below(Thesetransistorsarecalled pass transistors).



The connection of pass transistors in series will degrade the logic 1 level / into inverter 2 so that the output willnot be a properlogic 0 level. The criticalcondition is , whenpoint Ais at 0 volts and B is thus at VDD. but the voltage into inverter 2at point C is now reduced from VDD by the threshold voltage ofthe series pass transistor. With allpass transistor gates connected to VDDthere is a loss of Vtp, however manyare connected in series, since no static current flows throughthemand there can be no voltage drop inthe channels. Therefore, the input voltage to inverter 2 is

Vin2=VDD-Vtpwhere Vtp=thresholdvoltageforapasstransistor.

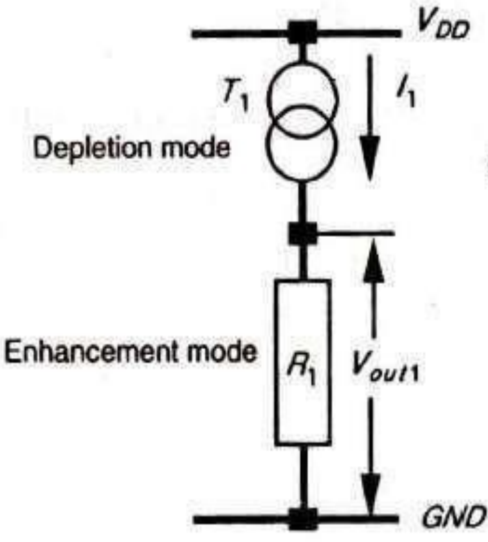
Let us consider the inverter 1 shown in Fig.(a) with input = VDD· If the input is at VDD , then the pull-down transistor T2 is conducting but with a low voltage across it; therefore, it is in its resistive regionrepresented byR1 inFig.(a) below. Meanwhile, the pullup transistorT1is insaturationand is represented as a current source.

Forthepulldowntransistor

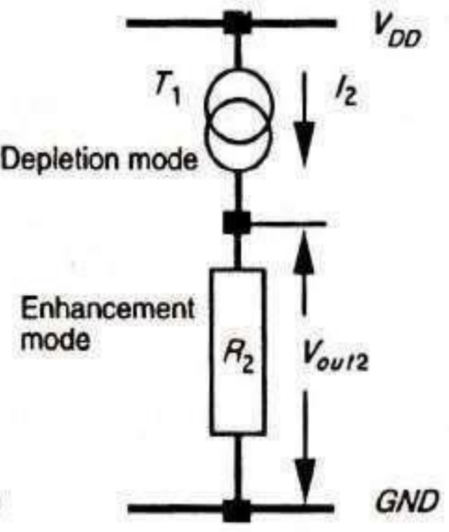
$$R_1 = \frac{V_{ds1}}{I_{ds}} = \frac{1}{K}\frac{L_{p.d.1}}{W_{p.d.1}}\left(\frac{1}{V_{DD} - V_t - \frac{V_{ds1}}{2}}\right)$$

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

SinceVdsissmall, Vds/2 canbeneglected intheaboveexpression.



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



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

So,

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

Now,fordepletion modepull-up transistorinsaturationwithVgs=0

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

Theproduct$1R_1$ = V_{out1}So,

$$V_{out1} = I_1 R_1 = \frac{Z_{p.d.1}}{Z_{p.u.1}} \left(\frac{1}{V_{DD} - V_t} \right) \frac{(V_{td})^2}{2}$$

Letusnowconsidertheinverter 2Fig.b.wheninput =V_{DD}-Vtp.

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

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

Whence,

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

Ifinverter 2istohavethesameoutputvoltageundertheseconditionsthen$V_{out1}=V_{out2}$.Thatis

$I_1R_1=I_2R_2$, therefore

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

Consideringthetypicalvalues

$$V_t = 0.2 V_{DD}$$

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

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

Therefore

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

Fromtheabovetheoryitisclearthat,for an n-MOStransistor

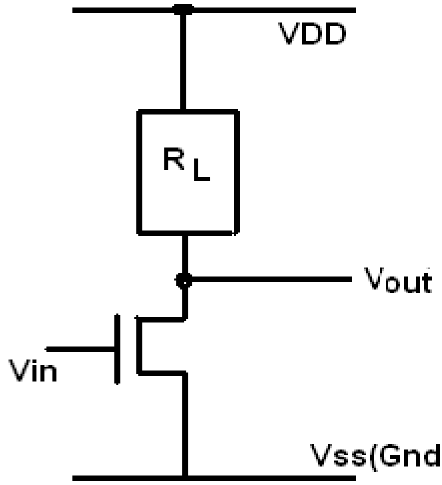
(i). An inverter driven directly from the output of another should have a $Z_{p.u}/Z_{pd.}$ ratio of ≥ 4/1.

(ii). AninverterdriventhroughoneormorepasstransistorsshouldhaveaZ$_{p.u.}$/Z$_{p.d}$ratioof≥8/1

ALTERMTIVEFORMSOFPULL–UP

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

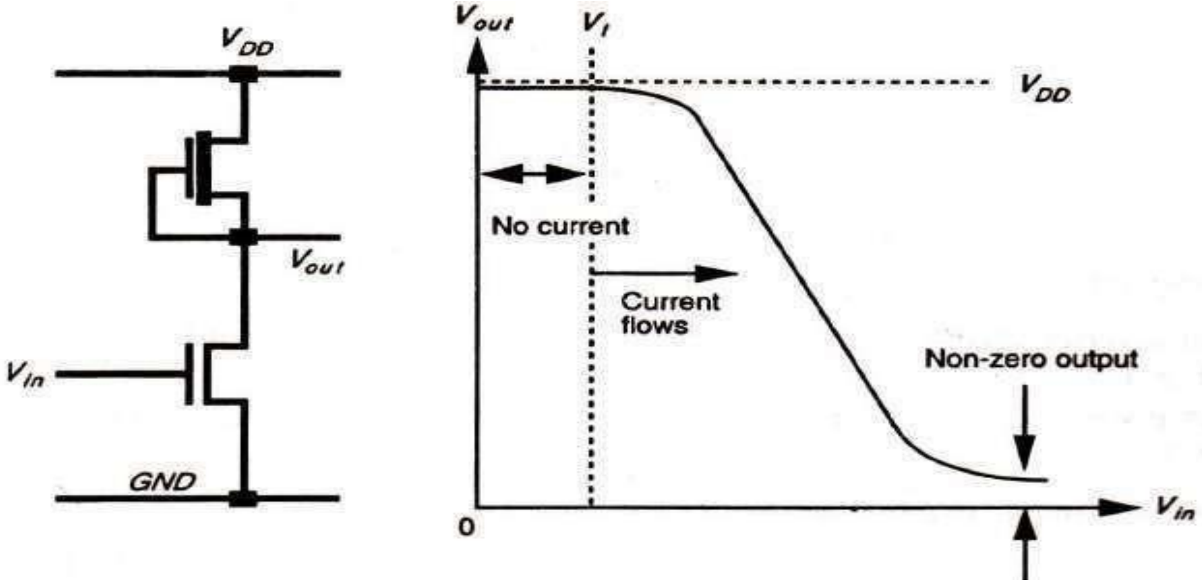
(i).Load resistance RL : This arrangement consists of a load resistor as apull-up as shown in the diagram below.But it is not widely used because of the large space requirements ofresistorsproduced in a silicon substrate.



nMOS depletion mode transistor pull-up : This arrangement consists of a depletion mode transistoraspull-up. Thearrangement andthetransfercharacteristicareshownbelow.Inthistype of arrangement we observe

(a) Dissipationis high,sincerailtorailcurrentflowswhenVin=logical1.

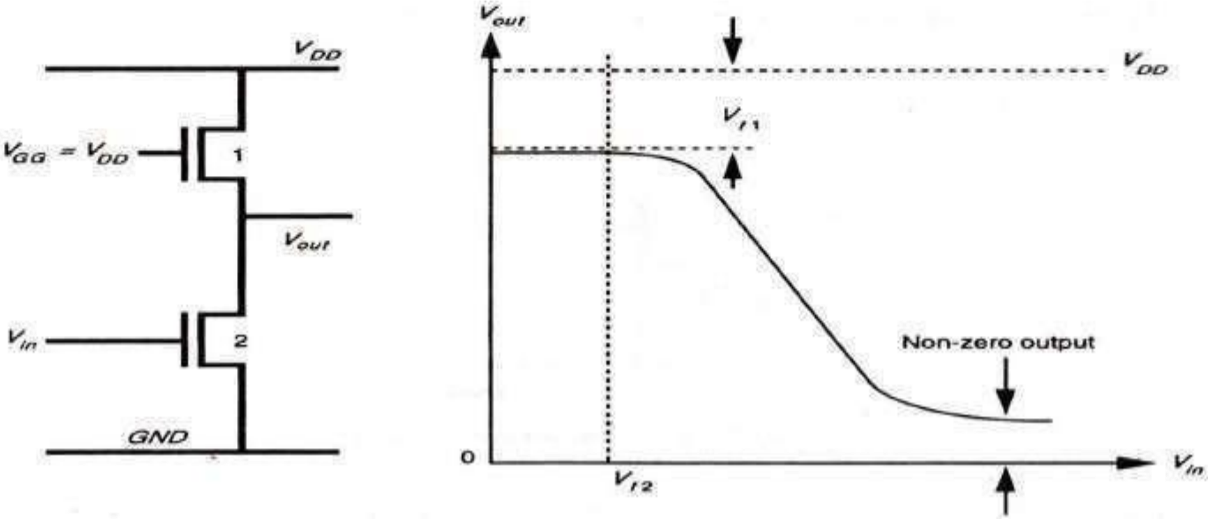
(b) Switchingofoutputfrom1to0begins whenVinexceedsVt, ofpull-downdevice.



nMOSdepletionmodetransistorpull-upandtransfercharacteristic

(c) Whenswitchingtheoutput from1to0,thepull-up device isnon-saturatedinitiallyandthis presents lower resistance through which to charge capacitive loads .

(ii)**nMOSenhancementmodepull-up:**Thisarrangementconsistsofan-MOSenhancement mode transistor as pull-up. The arrangement and the transfer characteristic are shownbelow.



nMOSenhancementmodepull-upandtransfercharacteristic

Theimportantfeaturesofthisarrangementare

(a) Dissipationis highsincecurrentflowswhenVin=logical1(V_{GG} isreturnedtoV_{DD}).

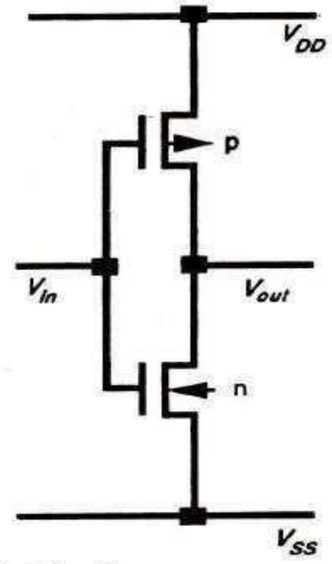
(b) Vout cannever reachV_{DD}(logicalI)ifV_{GG} =V_{DD}asisnormallythecase.

(c) V_{GG}maybederivedfromaswitchingsource,forexample,onephaseofaclock,sothat

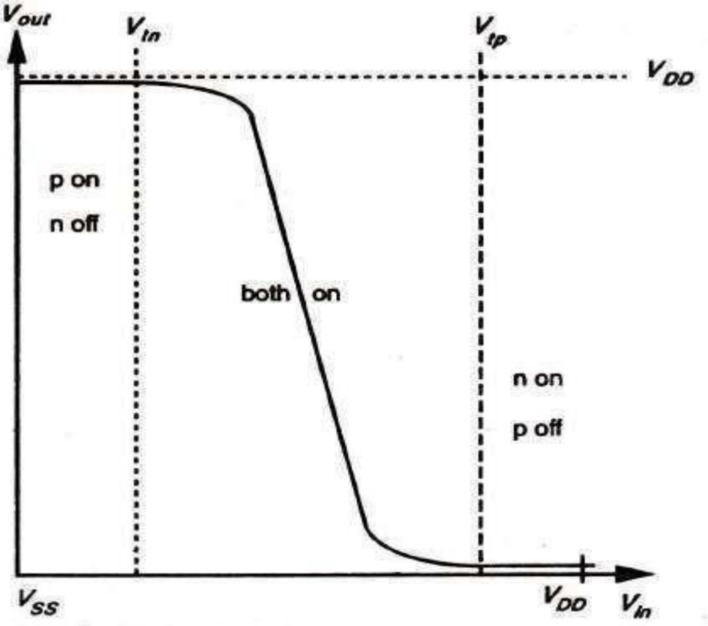
dissipationcanbegreatlyreduced.

(d) IfV_{GG}ishigherthanV_{DD}thenanextrasupplyrailisrequired.

(iii) **Complementarytransistorpull-up(CMOS):**ThisarrangementconsistsofaC-MOS arrangement as pull-up. The arrangement and the transfer characteristic are shownbelow



(a) Circuit



(b) Transfer characteristic

Thesalientfeaturesofthisarrangementare

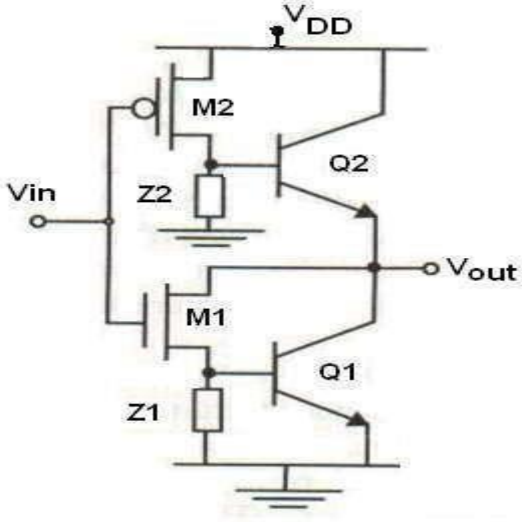
(a) Nocurrentflowseitherfor logical0orforlogical1inputs.

(b) Fulllogical1and0levelsarepresentedattheoutput.

(c) Fordevicesofsimilardimensionsthep-channelisslowerthanthe n-channeldevice.

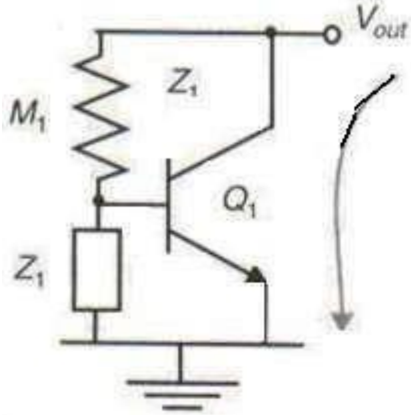
BiCMOS INVERTER:

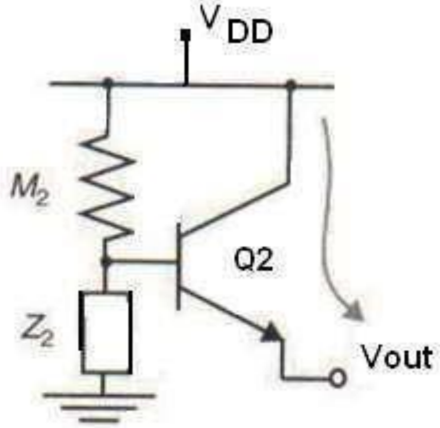
A BiCMOS inverter, consists of a PMOS and NMOS transistor (M2 and M1), two NPN bipolar junction transistors,(Q2 and Q1), and two impedances which act as loads(*Z*2 and *Z*1) as shown in the circuit below.



When input, Vin, is high (VDD), the NMOS transistor (M1), turns on, causing Q1 to conduct,while M2 and Q2 are off, as shown in figure (b) . Hence , a low (GND) voltage is translated to the output Vout. Onthe other hand, when the input is low, the M2 and Q2 turns on, while M1and Q1 turns off, resulting to a high output level at the output as shown in Fig.(b).

Insteady-stateoperation, Q1 and Q2 never turns onoroffsimultaneously, resulting to alower power consumption. This leads to a push-pull bipolar output stage. Transistors M1and M2, on the other hand, works as a phase-splitter, which results to ahigher input impedance.





The impedances Z2 and *Z*1 are used to bias the base-emitter junction of the bipolar transistor and to ensure that base charge is removed whenthe transistorsturnoff. For example whenthe input voltage makes a high-to-low transition, M1 turns off first. To turn off Q1, the base charge must be removed, which can be achieved by *Z*1.With this effect, transition time reduces. However, there exists a short time when both Q1 and Q2 are on, making a direct path from the supply (V_{DD}) to the ground. This results to a current spike that is large and has a detrimental effect on both the noise and power consumption, which makes the turning offof thebipolartransistorfast .

ComparisonofBiCMOSandC-MOStechnologies

The BiCMOS gates perform in the same manner as the CMOS inverter in terms of power consumption, because both gates display almost no static power consumption.

When comparing BiCMOS and CMOS in driving small capacitive loads, their performance are comparable, however, making BiCMOS consume more power than CMOS. On the other hand, driving larger capacitive loads makes BiCMOS in the advantage of consuming less power than CMOS, because the construction ofCMOS inverter chains are needed to drive large capacitance loads, which is not needed in BiCMOS.

The BiCMOS inverter exhibits a substantial speed advantage over CMOS inverters, especially when driving large capacitive loads. This is due to the bipolar transistor’s capabilityof effectively multiplying its current.

For very low capacitive loads, the CMOS gate is faster than its BiCMOS counterpartdue tosmall values of *Cint*. This makes BiCMOS ineffective when it comes to the implementation of internalgates for logic structures suchas ALUs, where associated load capacitances aresmall.

BiCMOS devices have speed degradation in the low supply voltage region and also BiCMOS is having greater manufacturing complexity than CMOS.

Assignment Questions:

1. Definethreshold voltage?DrivetheVtequationforMOStransistor.
2. ExplainwithneatdiagramsthevariousNMOSfabricationtechnology.
3. DrawandexplainBiCMOSinvertercircuit.
4. DiscusstheBasicElectricalPropertiesofMOSandBiCMOSCircuits.
5. Derivetheexpressionfor estimationofPull-UptoPull-Downratio ofann-MOS inverter driven by another n-MOS inverter.
6. DerivetherelationshipbetweenI$_{ds}$andV$_{ds}$
7. DerivetheexpressionfortransfercharacteristicsofCMOS Inverter.
8. WriteaboutBiCMOSfabrication inan-wellprocesswithadiagram.
9. DistinguishbetweenBipolar andCMOSdevicestechnologiesinbrief.
10. MentionabouttheBICMOSInvertersandalternativeBICMOSInverters.
11. Determinethepull-uptopulldownratio forNMOSinverterdrivenbyanotherNMOS Inverter
12. DrawthefabricationstepsofCMOStransistorandexplain itsoperationindetail.
13. DrawthefabricationstepsofNMOStransistorandexplain itsoperationindetail.

UNITII

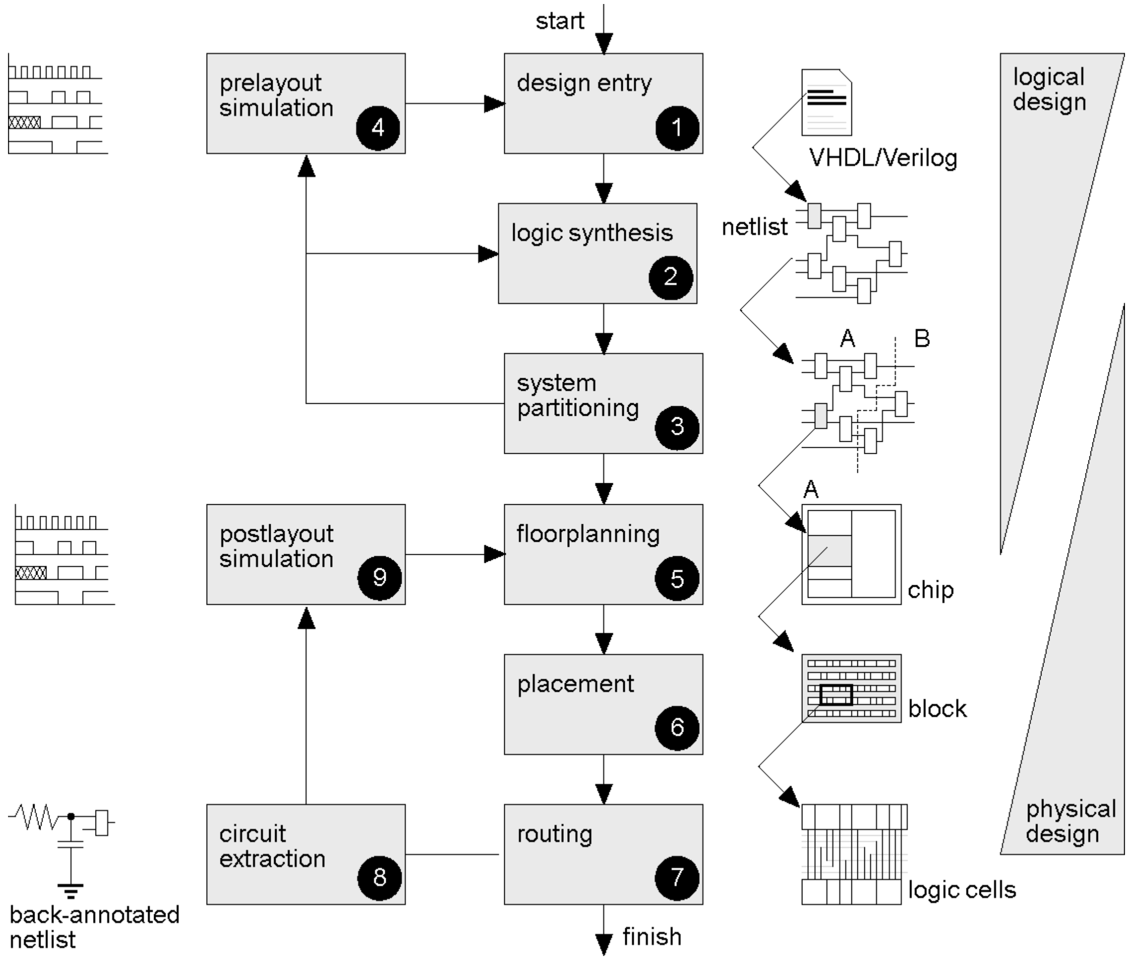
VLSICircuitDesignProcesses

- **VLSIDesignFlow**
- **MOSLayers**
- **StickDiagrams**
- **DesignRulesandLayout**
- **Lamda(λ)baseddesignrulesfor wires, contacts and Transistors**
- **LayoutDiagramsforNMOSand CMOS Inverters and Gates**
- **ScalingofMOScircuits**

VLSIDESIGNFLOW

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

Awell-tuneddesignflowcan helpdesignersgothroughthechip-creationprocessrelativelysmoothly and with a decent chance of error-free implementation. And, a skilful ICimplementation engineer canuse the design flow creativelytoshortenthe designcycle, resulting ina higher likelihood that the product will catch the market window.



Front-end design(Logicaldesign):

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

2. Logic synthesis – Generation of net list (logic cells and their connections) from HDL code. Logic synthesis consists of following steps: (i) Technology independent Logic optimization (ii) Translation: Converting Behavioraldescriptionto structuraldomain (iii) Technologymapping or Library binding

3. Systempartitioning-DividealargesystemintoASIC-sizedpieces

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

Back-enddesign(Physicaldesign):

5. Floorplanning-Arrangethe blocksofthe netlistonthechip

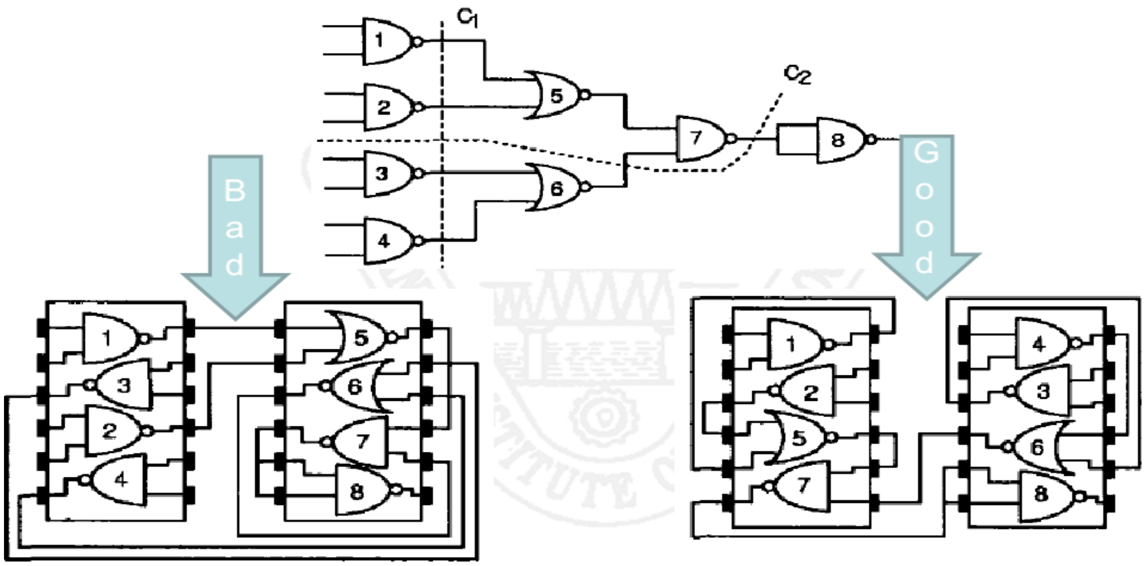
6. Placement-Decidethelocationsofcellsinablock

7. Routing-Maketheconnectionsbetweencellsandblocks

8. CircuitExtraction - Determinetheresistanceandcapacitanceoftheinterconnect

9. Post-layoutsimulation-Checkto seethedesignstillworkswiththeaddedloadsofthe interconnect

Partitioning



MOS LAYERS

MOSdesignisaimedatturningaspecificationintomasksforprocessingsilicontomeetthe specification. We have seenthat MOS circuits are formed on four basic layers

- N-diffusion
- P-diffusion
- PolySi
- Metal

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

STICKDIAGRAMS

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

"Astickdiagram isa cartoonofalayout."

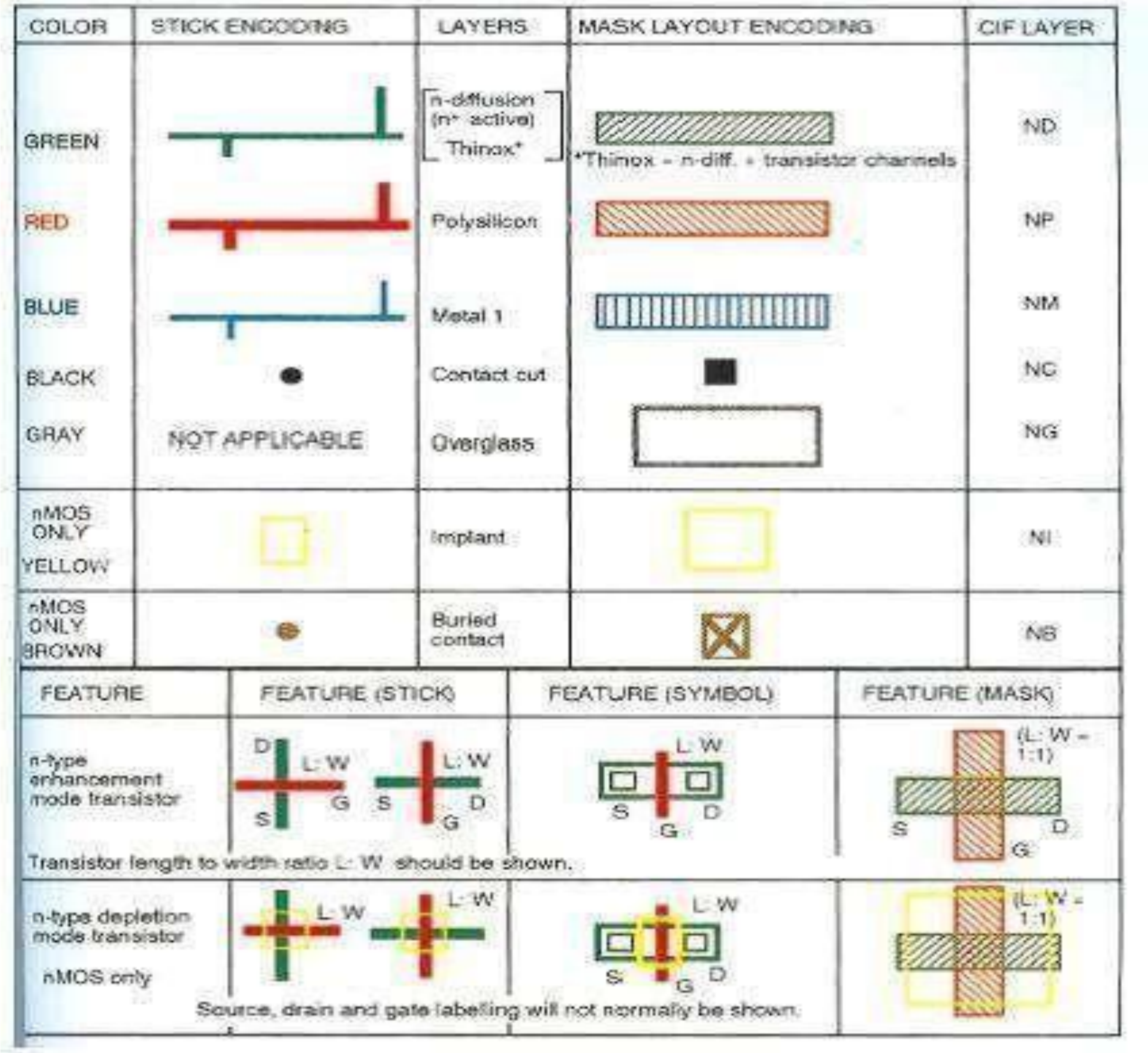
The designer draws a freehand sketch of a layout, using colored lines to represent the various process layers such as diffusion, metal and polysilicon. Where polysilicon crosses diffusion, transistors are created and where metalwires join diffusion or polysilicon, contacts are formed.

Forexample,inthecaseofnMOSdesign,

- Greencolor isusedfor n-diffusion
- Redforpolysilicon
- Blue formetal
- Yellowforimplant,andblackforcontactareas.

Monochromeencodingisalsousedinstickdiagramstorepresentthelayerinformation.

StickDiagrams–NMOSEncoding

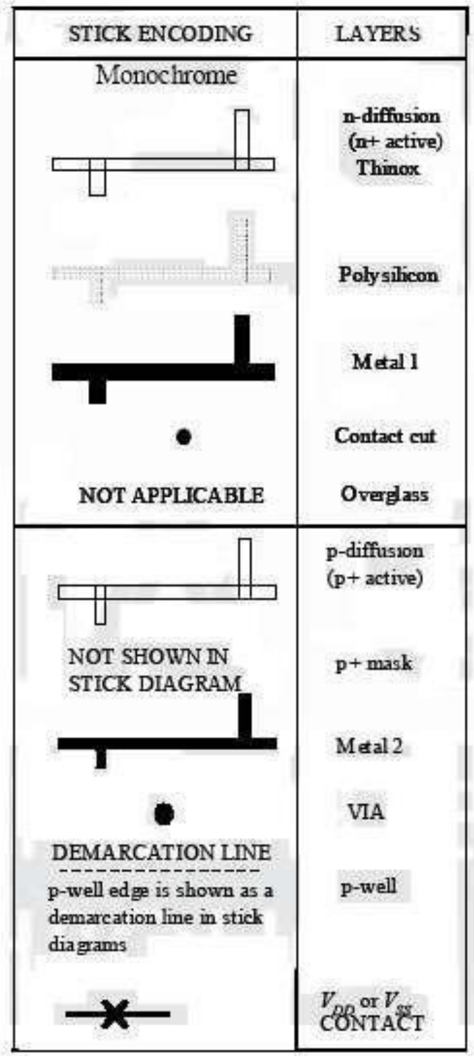


COLOR	STICK ENCODING	LAYERS	MASK LAYOUT ENCODING	CIF LAYER
GREEN		n-diffusion (n+ active) Thinox*	▨ *Thinox = n-diff. + transistor channels	ND
RED		Polysilicon	▨	NP
BLUE		Metal 1	▥	NM
BLACK	●	Contact cut	■	NC
GRAY	NOT APPLICABLE	Overglass	☐	NG
nMOS ONLY YELLOW	☐	Implant	☐	NI
nMOS ONLY BROWN	●	Buried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

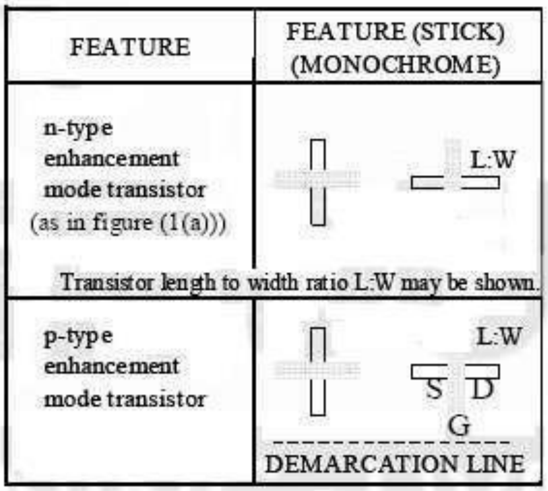
FEATURE	FEATURE (STICK)	FEATURE (SYMBOL)	FEATURE (MASK)
n-type enhancement mode transistor	D L:W G S L:W S D G	L:W S G D	(L:W = 1:1) S D G
Transistor length to width ratio L:W should be shown.			
n-type depletion mode transistor nMOS only	L:W L:W	L:W S G D	(L:W = 1:1)
Source, drain and gate labelling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

NMOSENCODING

CMOSENCODING



STICK ENCODING	LAYERS
Monochrome	
	n-diffusion (n+ active) Thinox
	Polysilicon
	Metal 1
●	Contact cut
NOT APPLICABLE	Overglass
	p-diffusion (p+ active)
NOT SHOWN IN STICK DIAGRAM	p+ mask
	Metal 2
●	VIA
DEMARCATION LINE - - - - - - - - - - - - - - - - p-well edge is shown as a demarcation line in stick diagrams	p-well
—✕—	V_{DD} or V_{SS} CONTACT

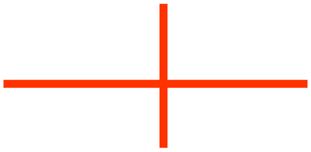


FEATURE	FEATURE (STICK) (MONOCHROME)
n-type enhancement mode transistor (as in figure (1(a)))	L:W
Transistor length to width ratio L:W may be shown.	
p-type enhancement mode transistor	L:W S D G - - - - - - - - - - - - - - - - DEMARCATION LINE

StickDiagrams–SomeRules

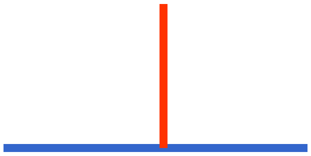
Rule1:

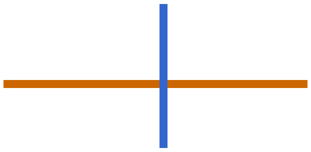
When two or more ‘sticks’ of the same type cross or toucheach other that represents electrical contact.

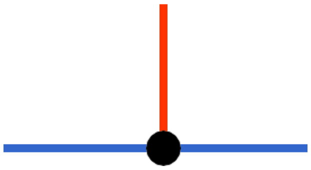


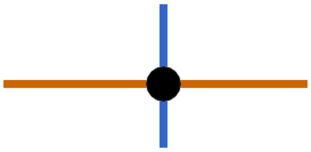
Rule2:

Whentwoormore“sticks”ofdifferent type crossortoucheachotherthereisno electrical contact. (Ifelectrical contact is needed we have to show the connection explicitly)









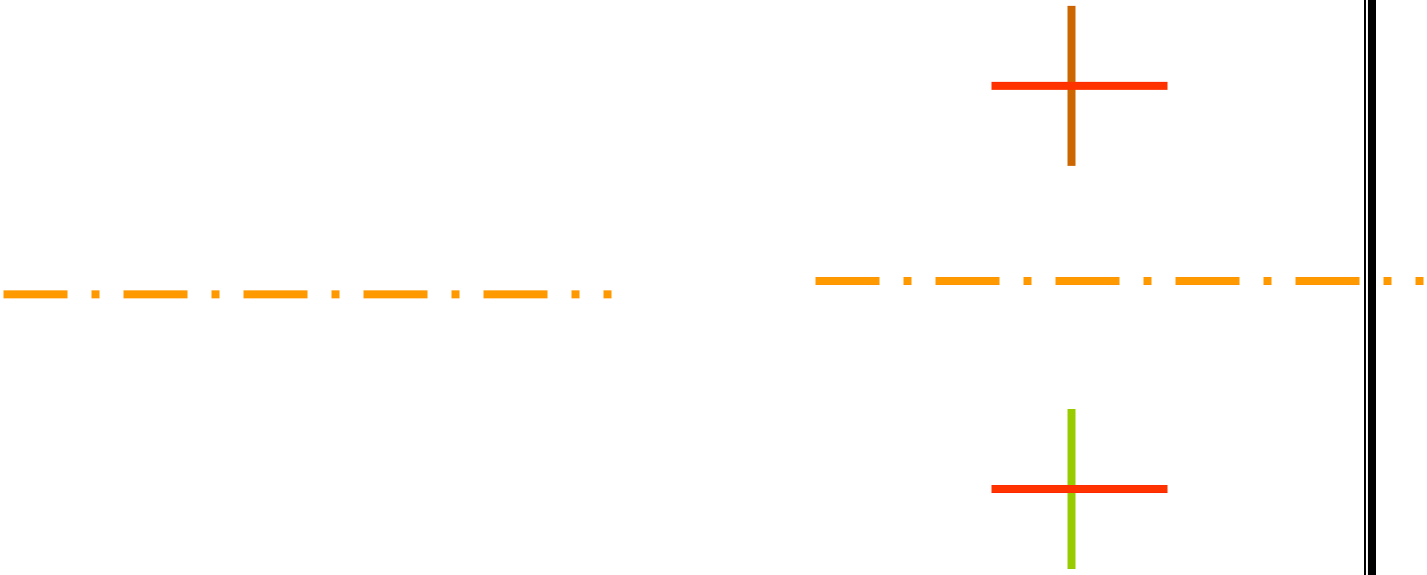
Rule 3:

When a poly crosses diffusion it represents a transistor.

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

Rule 4:

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



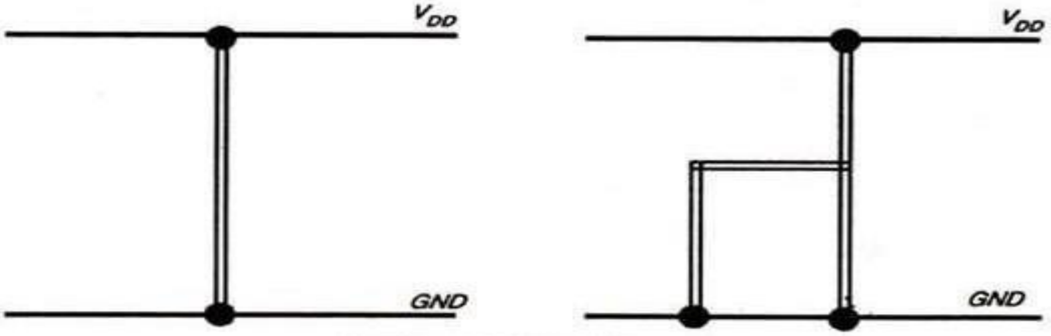
nMOSDesignStyle :

TounderstandthedesignrulesfornMOSdesignstyle,let usconsiderasingle metal,single polysilicon nMOS technology.

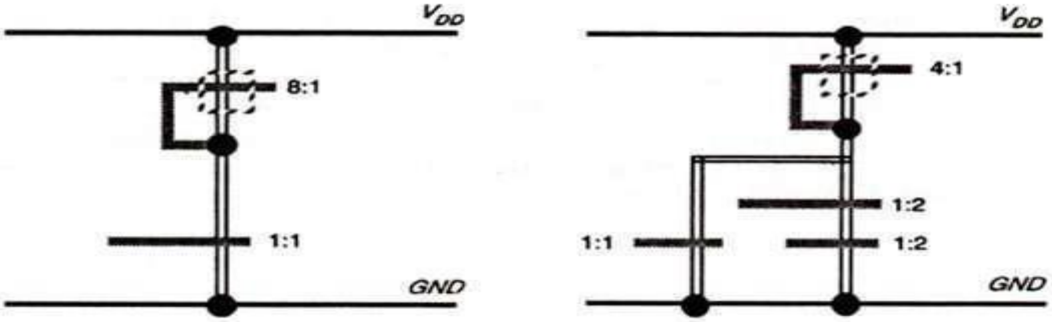
ThelayoutofnMOSis basedonthefollowingimportantfeatures.

- n-diffusion[n-diff.]andotherthinoxideregions[thinox] (green);
- polysilicon1[poly.]-sincethereisonlyonepolysilicon layerhere(red);
- metal1[metal]-sinceweuseonlyonemetallayerhere(blue);
- implant(yellow);
- contacts(blackor brown[buried]).

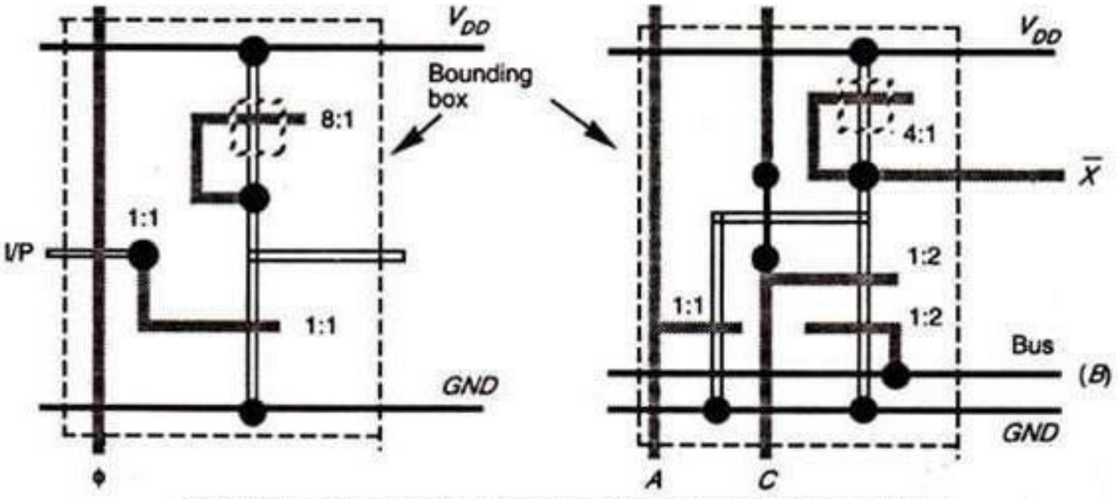
A transistor is formed wherever poly. crosses n-diff. (red over green) and all diffusion wires (interconnections) are n-type (green).When starting a layout, the first step normally taken is to draw the metal(blue) $V_{DD}$and GND rails in parallelallowing enoughspace betweenthem for the other circuit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



(a) Rails and thinox paths



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

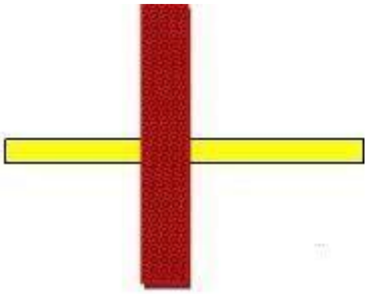


(c) Buses, control signals, interconnections, and 'leaf-cell' boundaries

CMOSDesignStyle:

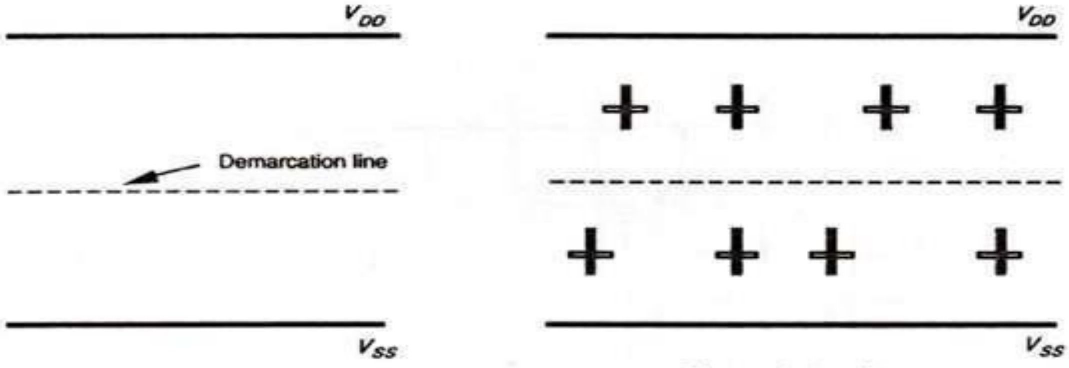
The CMOS designrules are almost similar and extensions of n-MOS designrules except the Implant (yellow) and the buried contact (brown). In CMOS design Yellow is used to identify p transistors and wires, as depletion mode devices are not utilized. The two types of transistors 'n' and 'p', are separated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

N-type (red over green)

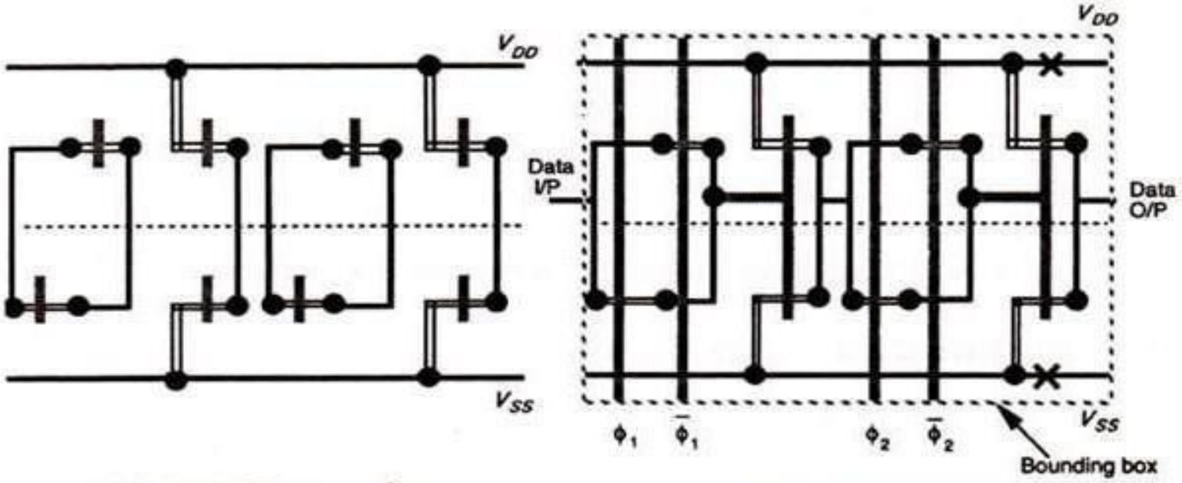


P-type (red over yellow)

Diffusion paths must not cross the demarcation line and n-diffusion and p-diffusion wiresmust not join. The 'n' and 'p' features are normally joined by metal where a connection is needed. Their geometry willappear when thestick diagramis translatedtoa masklayout. However, onemust not forget to place crosses on VDD and Vss rails to represent the substrate and p-well connection respectively. The design style is explained by taking the examplethedesign of a singlebit shift register. Thedesign begins with the drawing of the VDD and Vss rails in parallel and in metal and the creation of an (imaginary) demarcation line in-between, as shown in Fig.below. The n-transistors arethen placed belowthis lineand thus close to Vss, while p-transistors are placed above the line and below VDD In both cases, the transistors are conveniently placed with their diffusion paths parallel to the rails (horizontal in the diagram) as shown in Fig.(b). Asimilar approach can betaken with transistors in symbolicform.



(a) Rails and demarcation line (b) n- and p-transistors

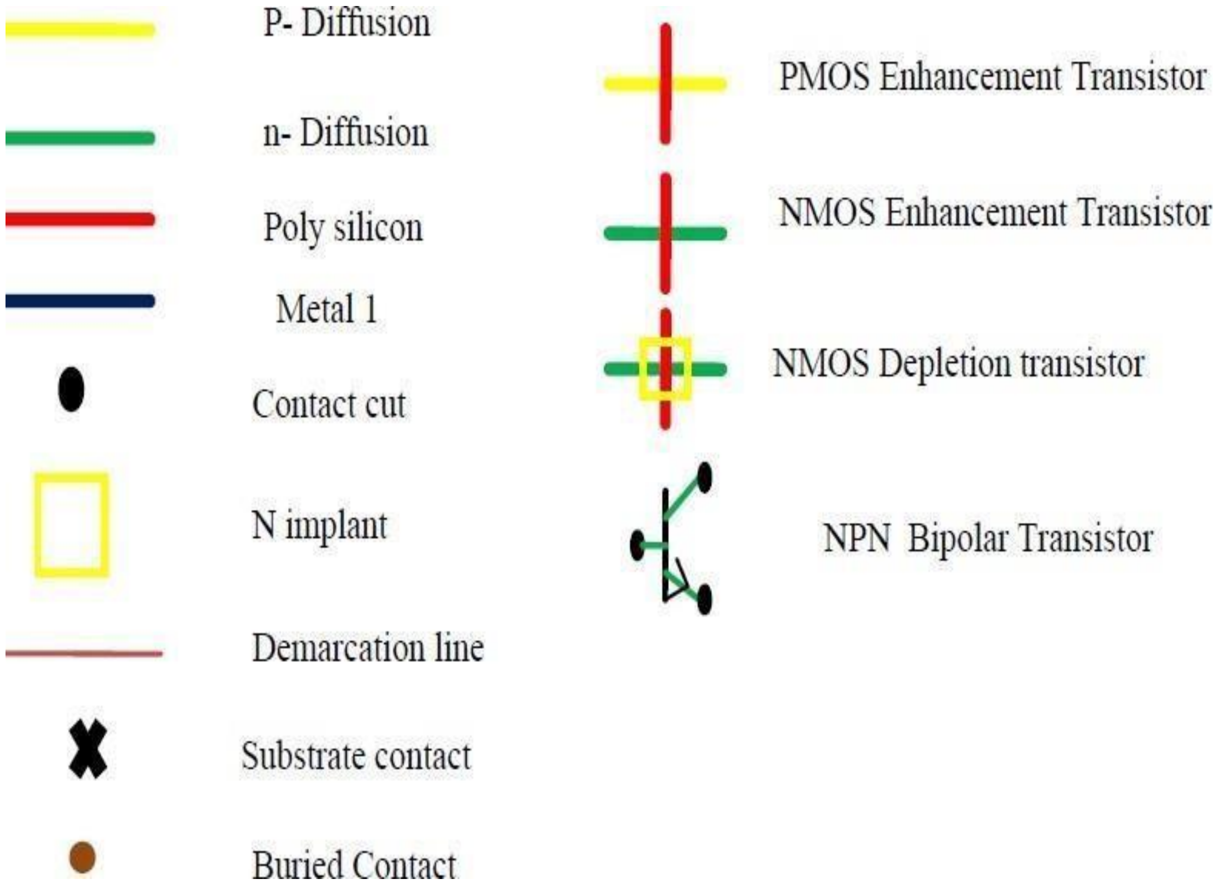


(c) Metal and diffusion connections (d) Remaining interconnections

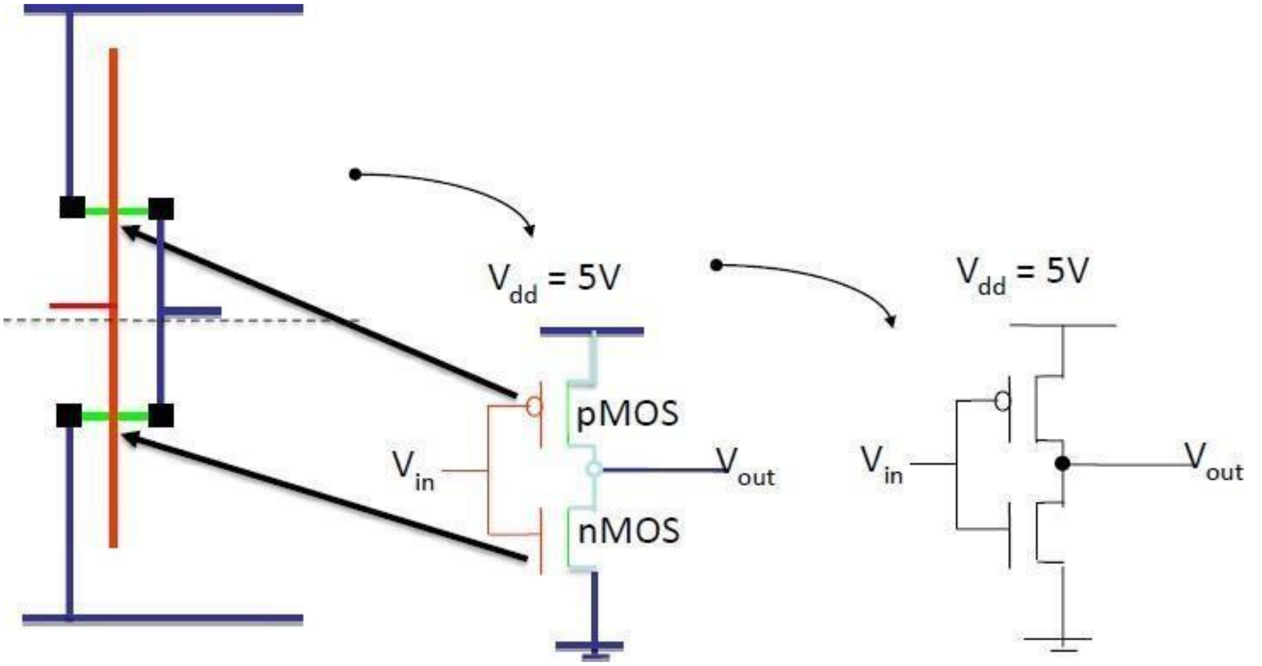
Fig.CMOSsticklayoutdesignstyle(a,b,c,d)

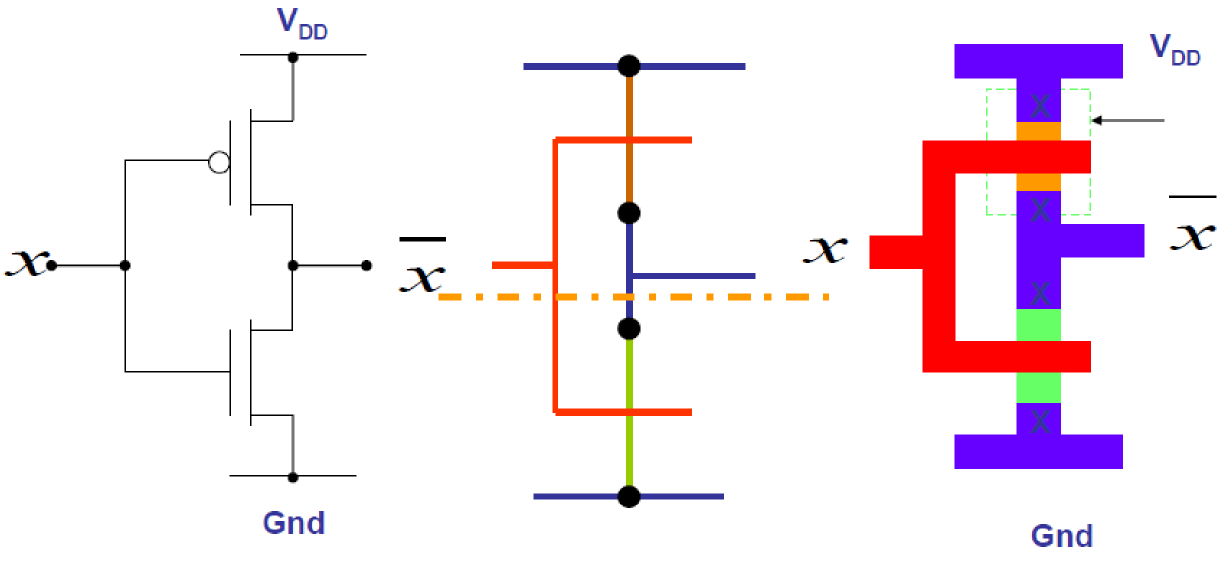
The n- along with the p-transistors are interconnected to the rails using the metal and connect as Shown in Fig.(d). It must be remembered that only metal and poly-silicon can cross the demarcation line but with that restriction, wires can run-in diffusion also. Finally, the remaining interconnections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

StickDiagrams:

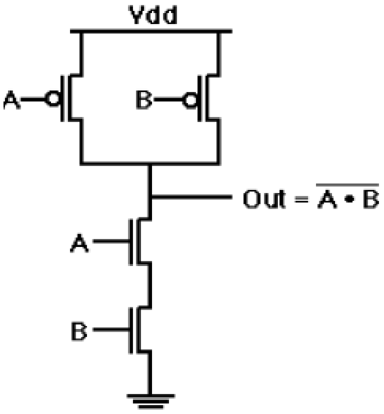


ExamplesofStickDiagramsCMOSInverter

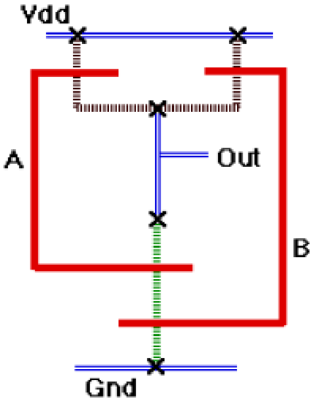




Contd….



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

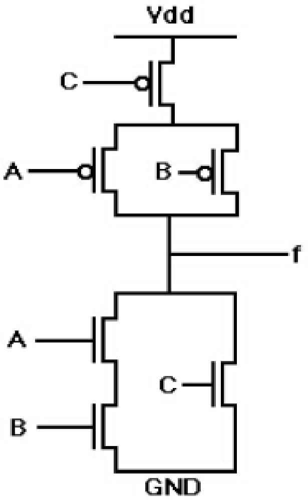


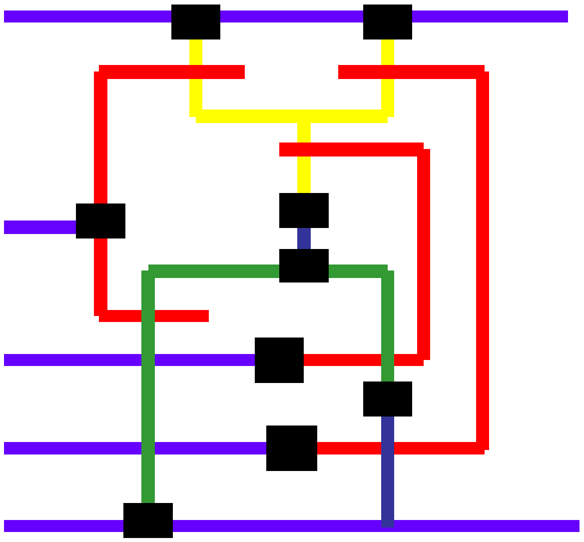
1. Pull-down: Connect to ground If A=1 AND B=1

2. Pull-up: Connect to Vdd If A=0 OR B=0

Fig.CMOSNANDgate

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





DesignRulesandLayout

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

- Photoresistshrinkage,tearing.
- Variationsinmaterialdeposition, temperatureandoxidethickness.
- Impurities.
- Variations across a wafer.

These leadtovariousproblems like:

- **Transistorproblems:**

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

- **Wiringproblems:**

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

- **Oxide problems:**

  Variationsinheight.

  Lack of planarity.

- **Viaproblems:**

  Viamaynotbecutallthewaythrough.

Undersizeviahastoomuchresistance.

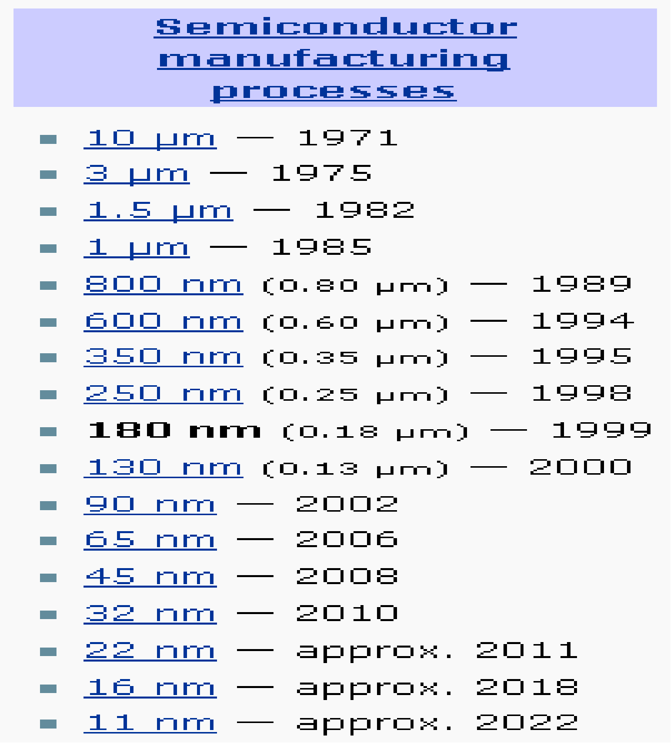
Via maybe too large and create short.

To reduce these problems, the design rules specify to the designer certain geometric constraints on the layout artwork so that the patterns on the processed wafers will preserve the topologyand geometryof the designs. This consists of minimum-width and minimum-spacing constraints and requirementsbetweenobjectsonthe same ordifferent layers. Apart fromfollowing a definite set of rules, design rules also come by experience.

Whyweusedesignrules?

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

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



Semiconductor manufacturing processes

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

Designrulesdefinerangesforfeatures

Examples:

- min. wirewidthstoavoid breaks
- min.spacingto avoidshorts
- minimumoverlapstoensurecompleteoverlaps
  – Measuredinmicrons
  – Requiredforresolution/tolerancesofmasks

Fabrication processes defined by minimum channel width

– Alsominimumwidthofpolytraces

– Defines“howfast”afabricationprocessis

TypesofDesignRules

Thedesignrulesprimaryaddresstwoissues:

1. Thegeometricalreproductionoffeaturesthat canbe reproducedbythe maskmakingand lithographical process, and
2. The interactionbetweendifferentlayers.

Thereareprimarilytwoapproaches indescribingthedesignrules.

1. Linearscalingispossibleonlyoveralimitedrangeofdimensions.
2. Scalabledesignrulesareconservative.Thisresults inoverdimensionedand lessdense design.
3. Thisruleisnotusedinreal life.

1. ScalableDesign Rules(e.g.SCMOS,λ-baseddesignrules):

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

Thekeydisadvantages ofthis approachare:

2. AbsoluteDesignRules(e.g. µ-baseddesignrules):

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

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

LAMBDA-BASED DESIGN RULES:-

- *Lambda-based*(scalableCMOS)designrulesdefinescalablerulesbasedonλ (which is half of the minimum channel length)
  - classesofMOSISSCMOSrules:SUBMICRON,DEEPSUBMICRON
- Stickdiagramisadraftofreallayout,it servesasanabstract viewbetweenthe schematic and layout.
- Circuitdesigneringeneralwanttighter,smallerlayoutsforimprovedperformance and decreased silicon area.
- Ontheotherhand,theprocessengineerwantsdesignrulesthat result ina controllable and reproducible process.
- Generallywefindtherehastobeacompromise foracompetitivecircuit to be produced at a reasonable cost.
- Allwidths,spacing,anddistancesarewrittenintheform
- λ=0.5Xminimumdrawntransistorlength
- Designrulesbased onsingleparameter,λ
- Simpleforthedesigner
- Wideacceptance
- Providefeaturesize independent way of settingoutmask
- Ifdesignrulesareobeyed,maskswillproduce working circuits
- Minimumfeaturesizeisdefinedas2λ
- Usedtopreservetopologicalfeaturesonachip
- Preventsshorting,opens,contactsfromslipping outofareatobecontacted

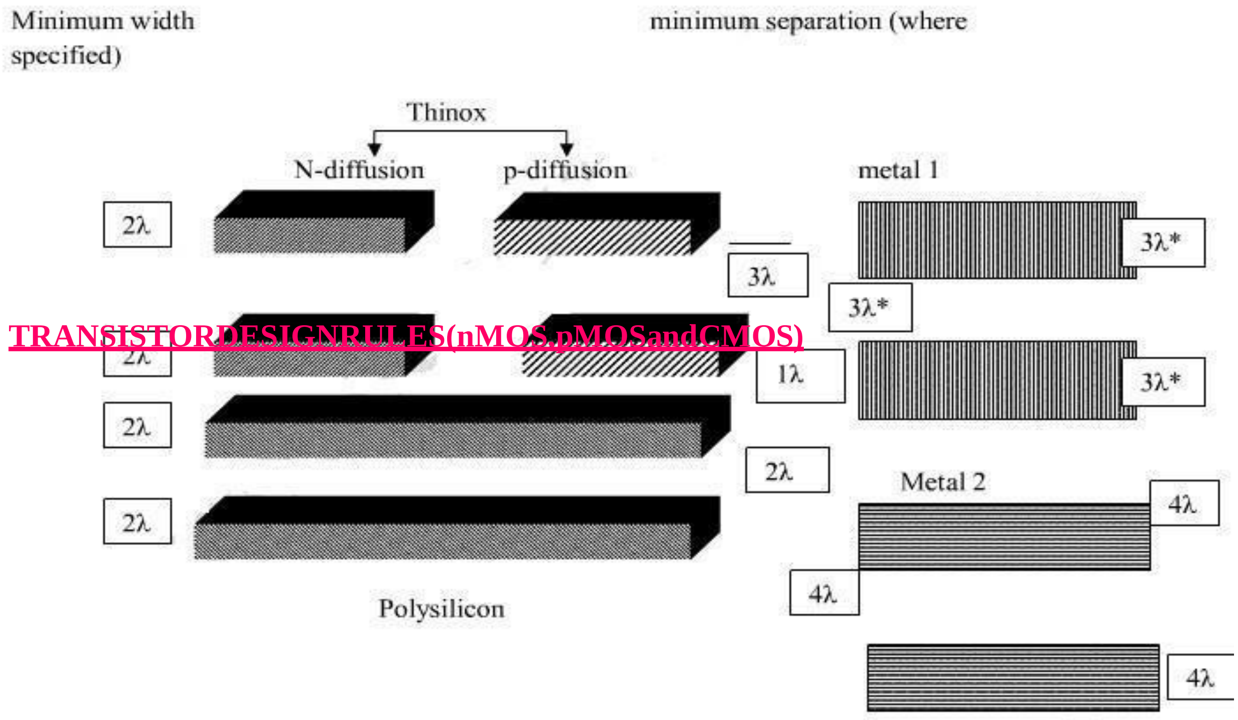
LAMBDA BSED RULES

MINIMUM WIDTH AND SPACING RULES

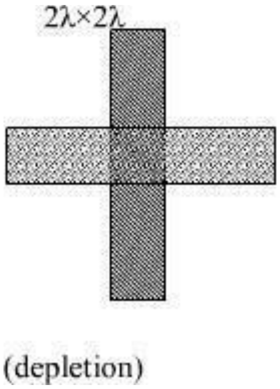
LAYER	TYPE OF RULE	VALUE
POLY	Minimum Width Minimum Spacing	2λ 2λ
N/P DIFFUSION	Minimum Width Minimum Spacing	3λ 3λ
N-WELL	Minimum Width Minimum Spacing	3λ 3λ
P-WELL	Minimum Width Minimum Spacing	3λ 3λ
METAL1	Minimum Width Minimum Spacing	3λ 3λ

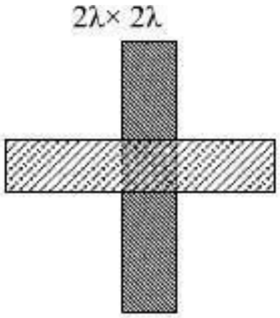
DESIGNRULESFORWIRES(nMOSandCMOS)

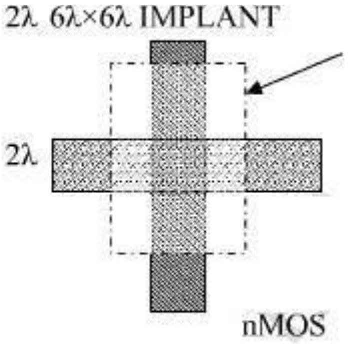
Design rules and layout methodology based on the concept of λ provide a process and feature size independent way of setting out mask dimensions to scale. All paths in layers are dimensioned in λ units and subsequently λ can be allocated an appropriate value compatible with the feature size of the fabrication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



TRANSISTORDESIGNRULES(nMOS,pMOSandCMOS)





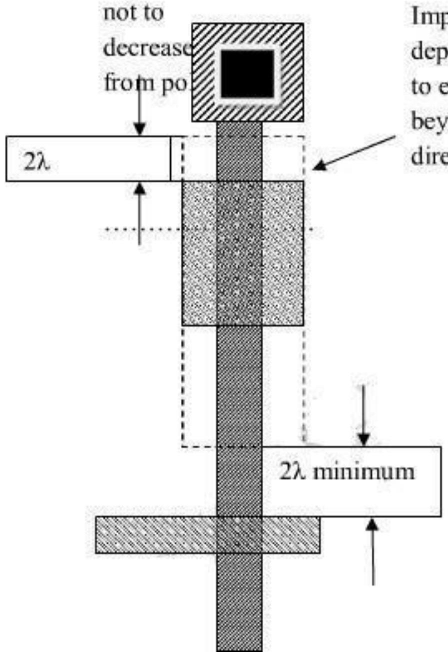


n MOS (enhancement) pMOS (enhancement)

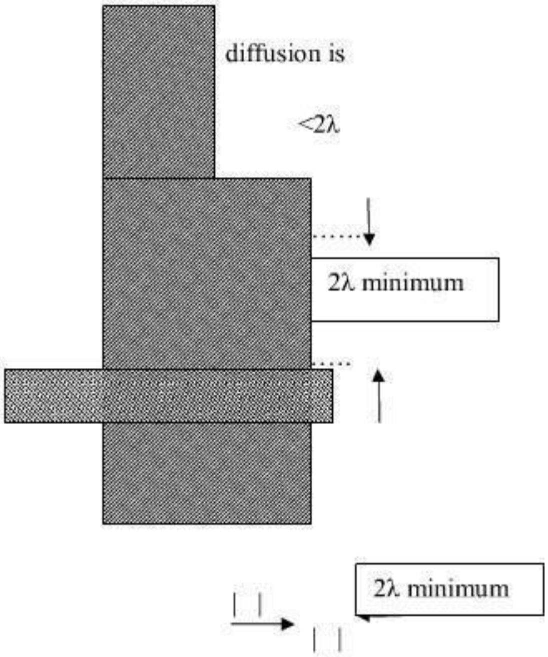
Separation from contact cut to transistor

not to decrease from po

Implant for an nMOS depletion mode transistor to expend 2λ minimum beyond channel in all directions



diffusion is <2λ



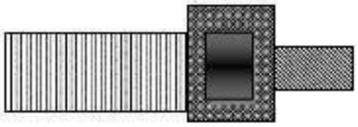
polysilicon to extend a minimum of 2λ beyond diffusion boundaries (width constant)

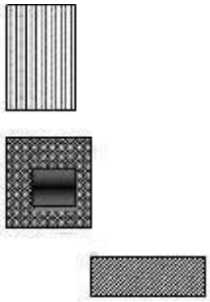
Separation from implant to another transistor

Key: ▩ polysilicon ▩ n-diffusion ▨ p-diffusion ▦ transistor channel ……………………………………………………………………………

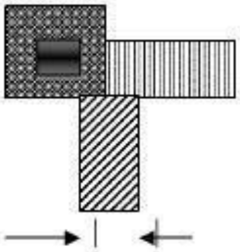
(Polysilicon over thinox)

metal 1 to polysilicon or to diffusion



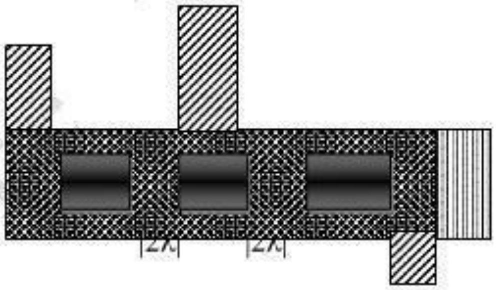


3λ minimum



cuts

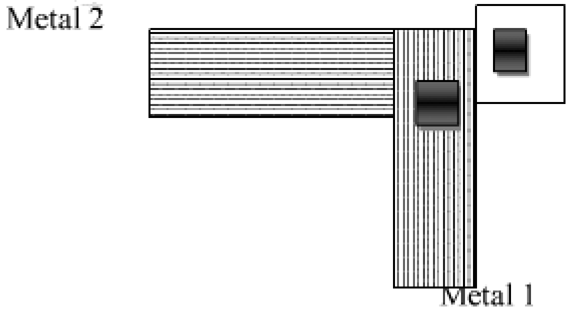
2λ minimum



minimum separation multiple

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

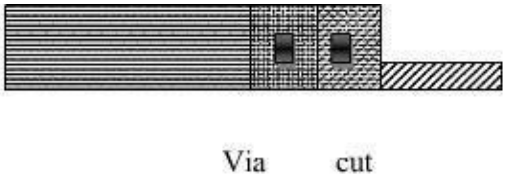
2λ minimum separation (if other spacing's allowed)



4λ×4λ area of overlap with

2λ×2λ via at center

Via and cut used to connect metal 2 to diffusion



Contacts (nMOS and CMOS)

CONTACTCUTS

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

InCMOSdesigns,poly.to diff.contactsarealwaysmadeviametal.Asimpleprocessis followed for making connections between metaland either oftheother two layers (as inFig.a), The 2λ. x 2λ. contact cut indicates an area in which the oxide is to be removed down to the underlying polysilicon or diffusion surface. When deposition of the metal layer takes place the metalisdepositedthroughthecontact cut areasontotheunderlying areaso that contact is made between the layers.

Theprocessismorecomplexforconnectingdiffusionto poly-siliconusingthebutting contact approach(Fig.b), Ineffect, a2λ. x2λcontact cut is madedownto eachofthe layersto be joined. The layers are butted together in such a waythat these two contact cuts become contiguous. Since the poly-silicon and diffusion outlines overlap and thin oxide under poly silicon acts as a mask in the diffusion process, the poly-silicon and diffusion layers are also buttedtogether.Thecontactbetweenthetwo butting layers isthenmade bya metaloverlayas shown in the Fig.

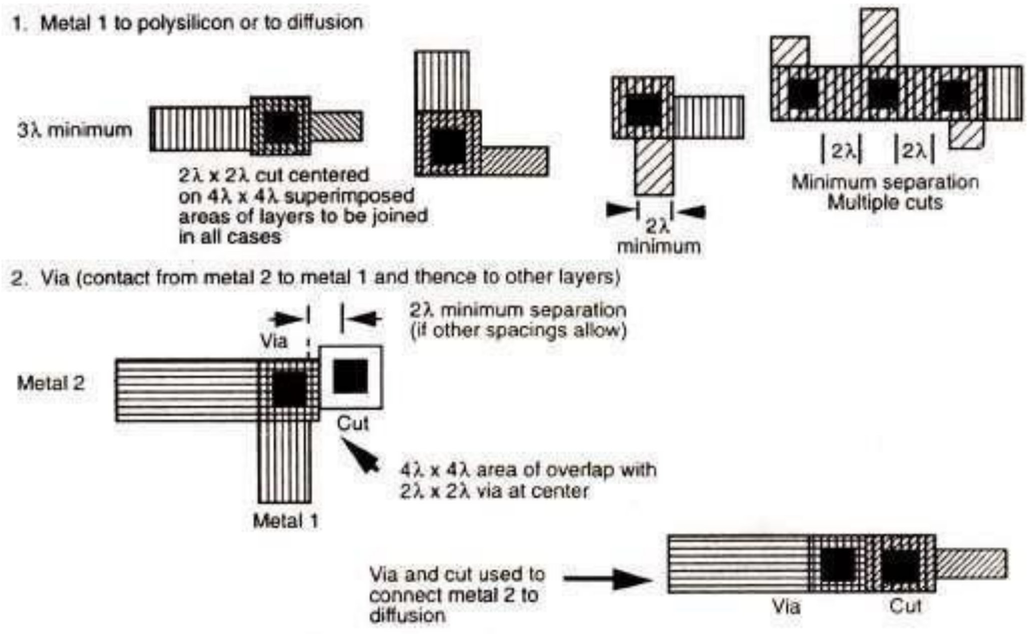


Fig.(a).n-MOS &C-MOSContacts

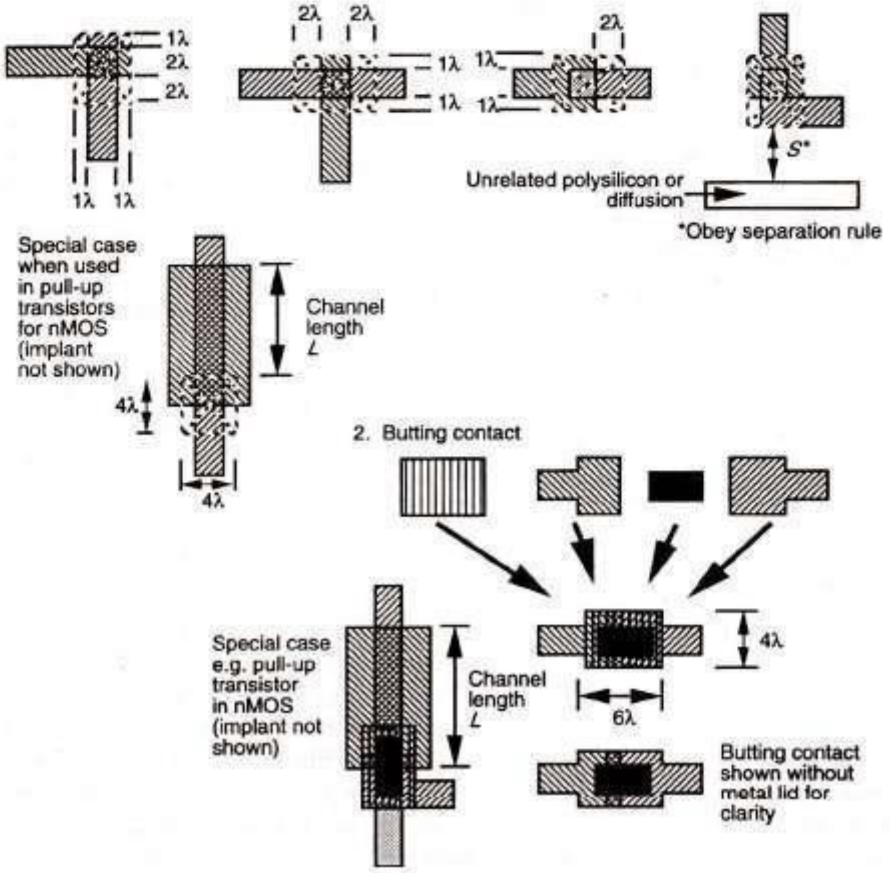


Fig.(b). Contactspoly-silicontodiffusion

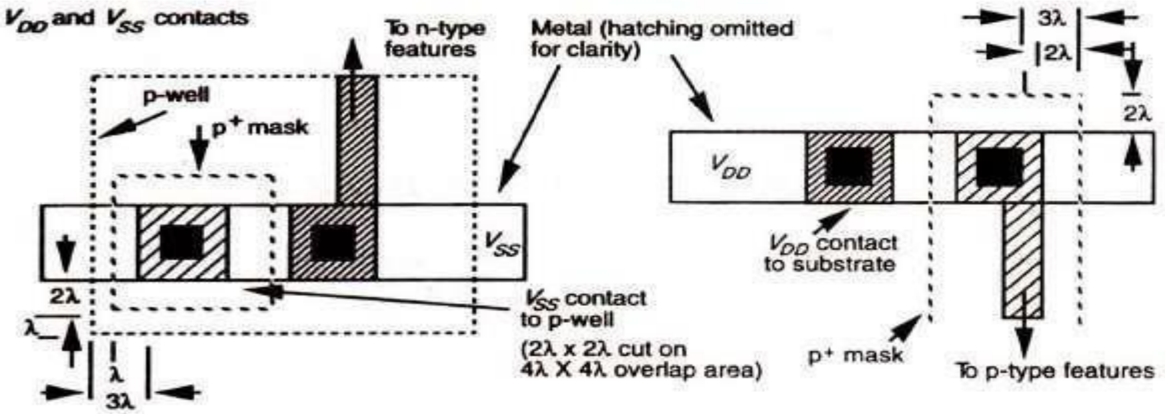
Inburiedcontact basically, layersare joinedovera2λ. x2λ. areawiththe buried contact cut extending by 1λ, in all directions around the contact area except that the contact cut extension is increased to 2λ. in diffusion paths leaving the contact area. This helps to avoid the formation of unwanted transistors. So this buried contact approach is simpler when compared to others. The, poly-silicon is deposited directly on the underlying crystalline wafer.When diffusion takes place, impurities will diffuse into the poly-silicon as well as into the diffusion region within the contact area. Thus a satisfactory connection between poly-silicon and diffusion is ensured. Buried contacts can be smaller in area than their butting contact counterparts and, since theyuse no metallayer, theyare subject to fewer designrule restrictions ina layout.

Otherdesignrules

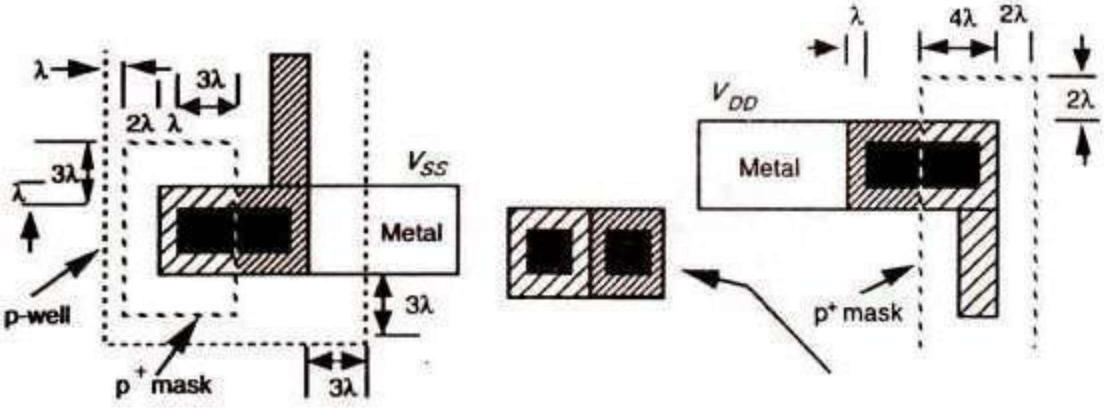
- DoubleMetalMOSprocess Rules
- CMOS fabricationismuchmorecomplexthannMOSfabrication
- 2umDoublemetal,Doublepoly.CMOS/BiCMOSRules
- 1.2umDoubleMetalsingle poly.CMOSrules

CMOSLambda-basedDesignRules:

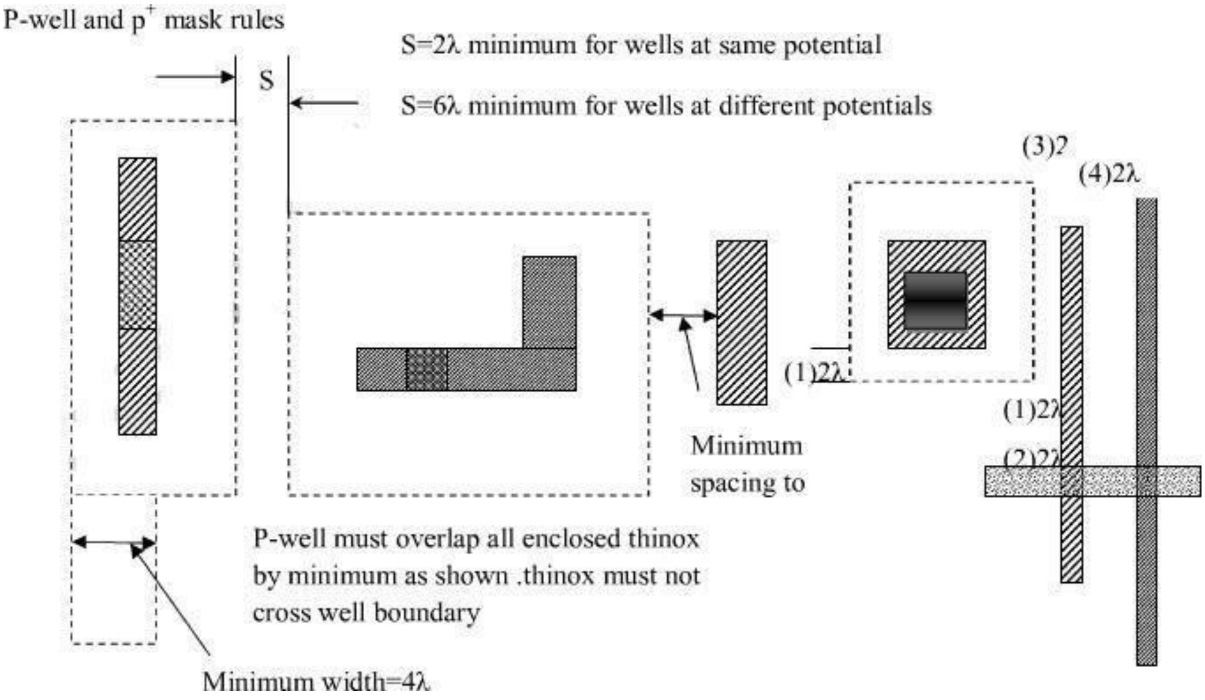
The CMOS fabrication process is more complex than nMOS fabrication. In a CMOS process, there are nearly 100 actual set of industrial design rules. The additional rules are concerned with those features unique to p-well CMOS, such as the p-well and p+ mask and the special 'substrate contacts. The p-well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



Inthediagramaboveeachofthearrangementscanbemergedintosinglesplitcontacts.



Fromtheabovediagramitisalsoclear thatsplitcontactsmayalsobemadewithseparatecuts.



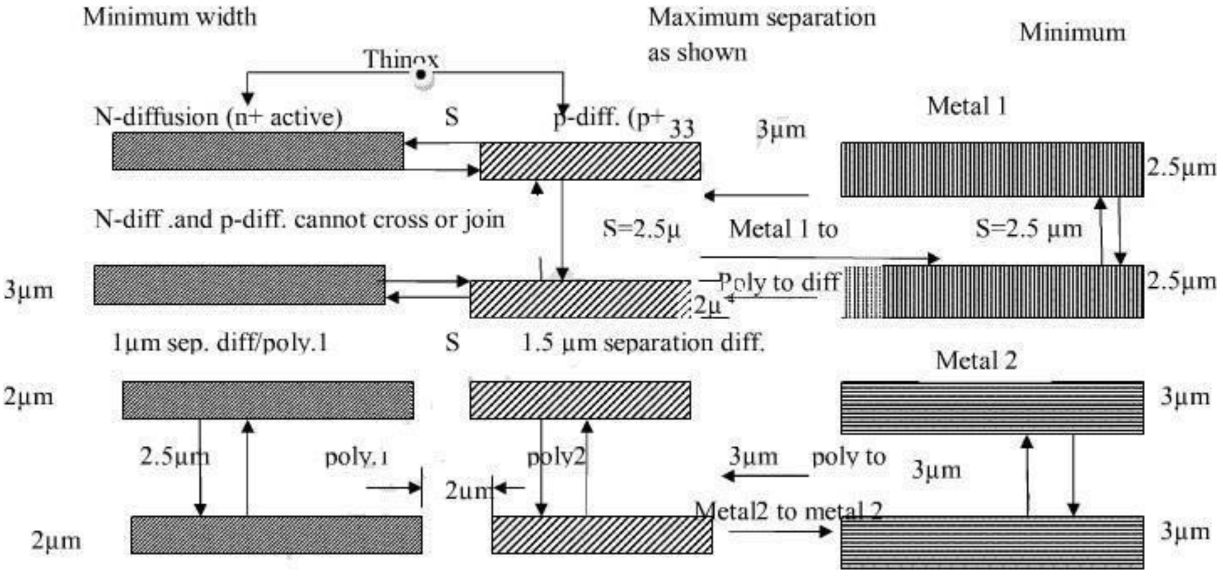
The CMOS rules are designed based onthe extensions ofthe Mead and Conway concepts and also by excluding the butting and buried contacts the new rules for CMOS design are formed. These rules for CMOS design are implemented inthe above diagrams.

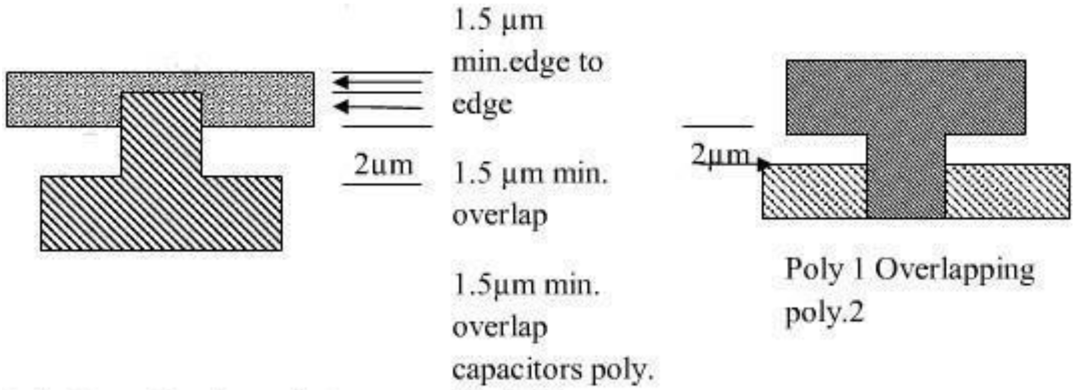
µMCMOSDesignrules

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

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

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





Poly 1 Overlapping poly.2

Poly. 2 overlapping poly.1

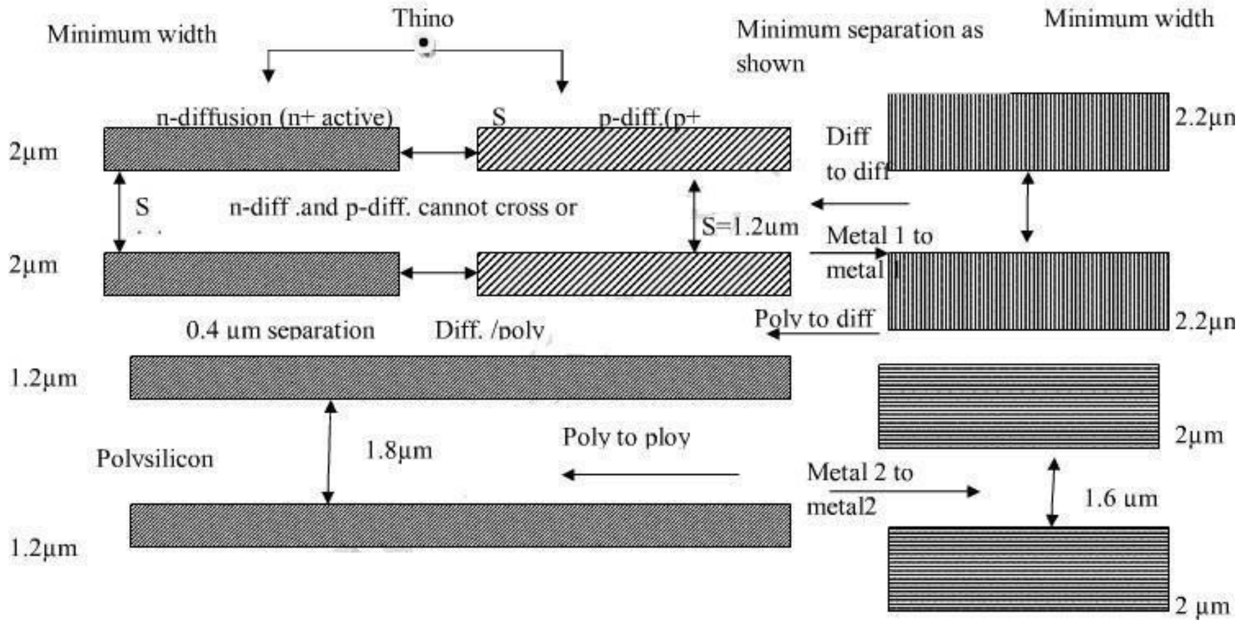
AVOID COINCIDENT EDGES WHERE METAL 1
AND METAL 2 RUNS FOLLOW THE SAME PATH
FOR >25µm LENGTH (UNDER LAP METAL 1

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

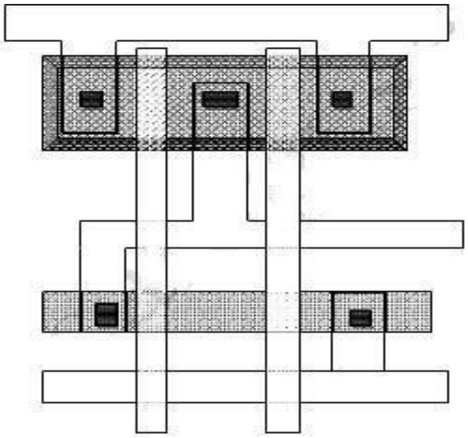
2µm DOUBLE METAL, SINGLE POLY CMOS RULES

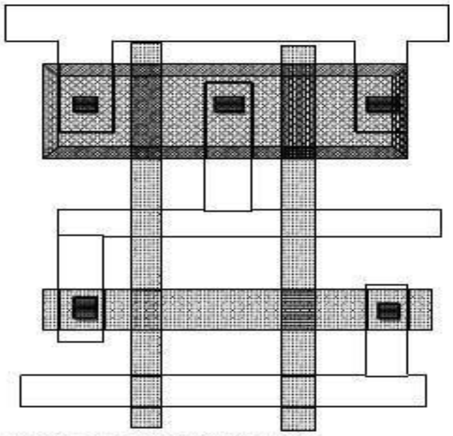
The orbit ^{TM} 1.2µm rules provide improved feature size. A separate set of micro based design rules accompany them

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

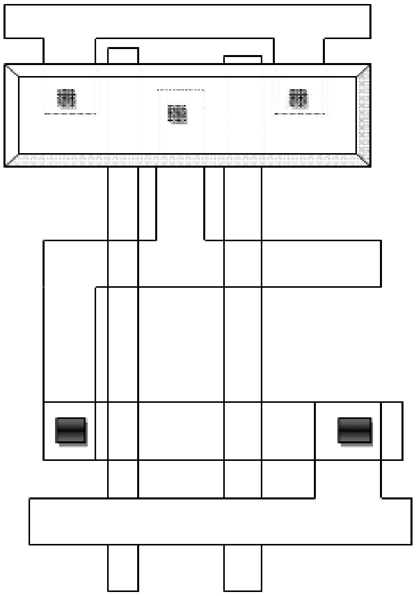


Avoid coincident edges where metal 1 and metal2 runs follow the same path for>25µm length (under lap metal 1 edges by 0.8 µm).

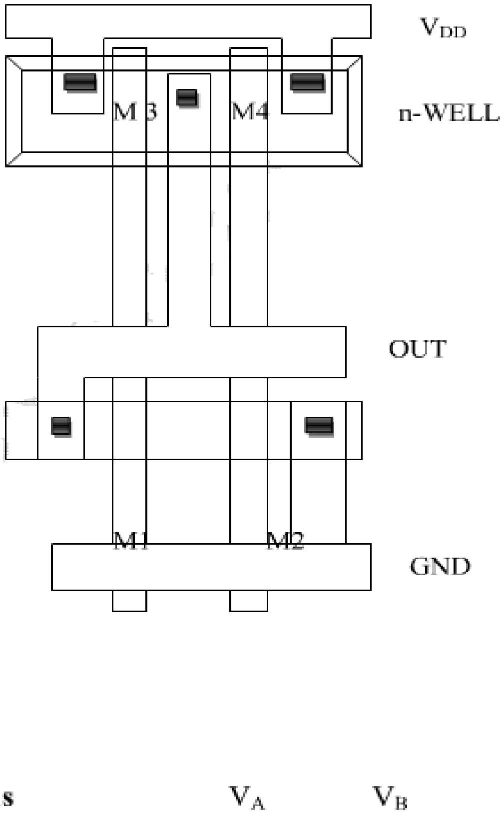




N-WELL AND ACTIVE AREA MASKs POLY MASK -> DEFINE NMOS AND . .. ……………………………………………………………………………PMOS TRANSISTORS

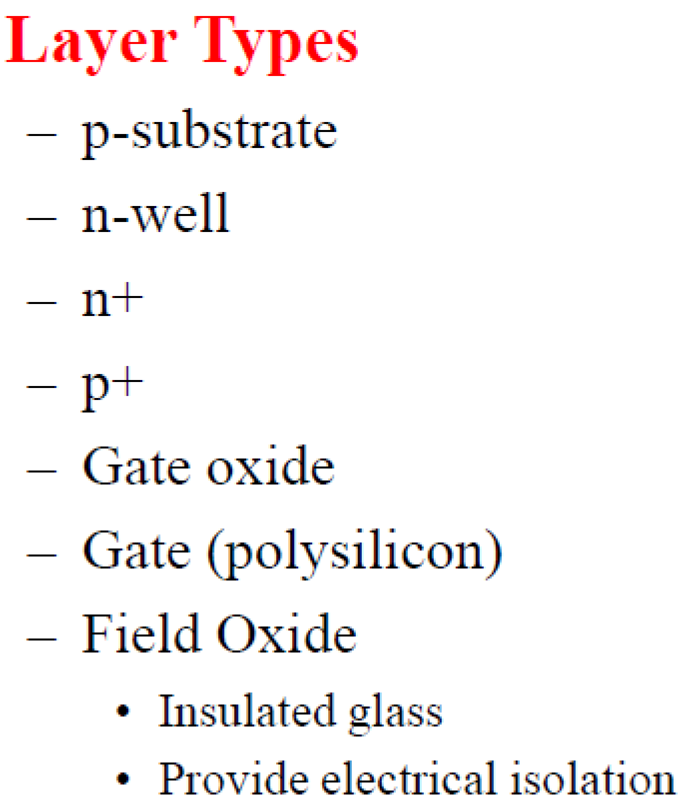


Metal mask for V_{DD}, GND and output connections



METAL –DIFFUSION CONSTANT MASK

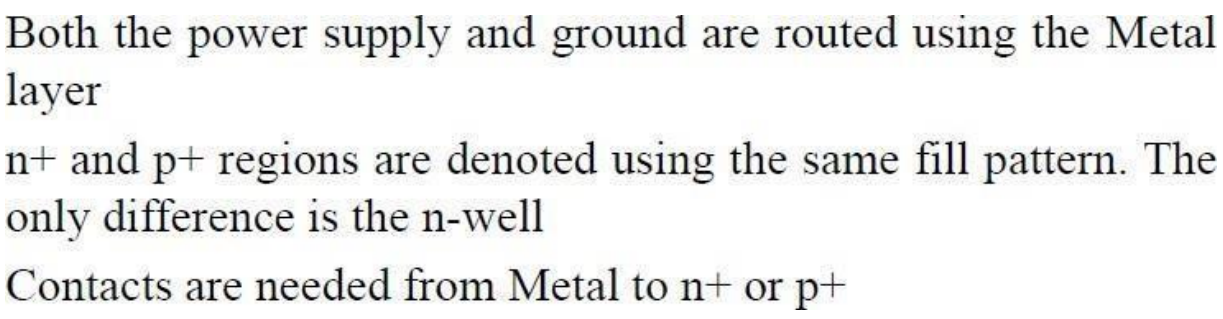
LayoutDiagramsforNMOSandCMOSInvertersandGates



Layer Types

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

BasicGateDesign

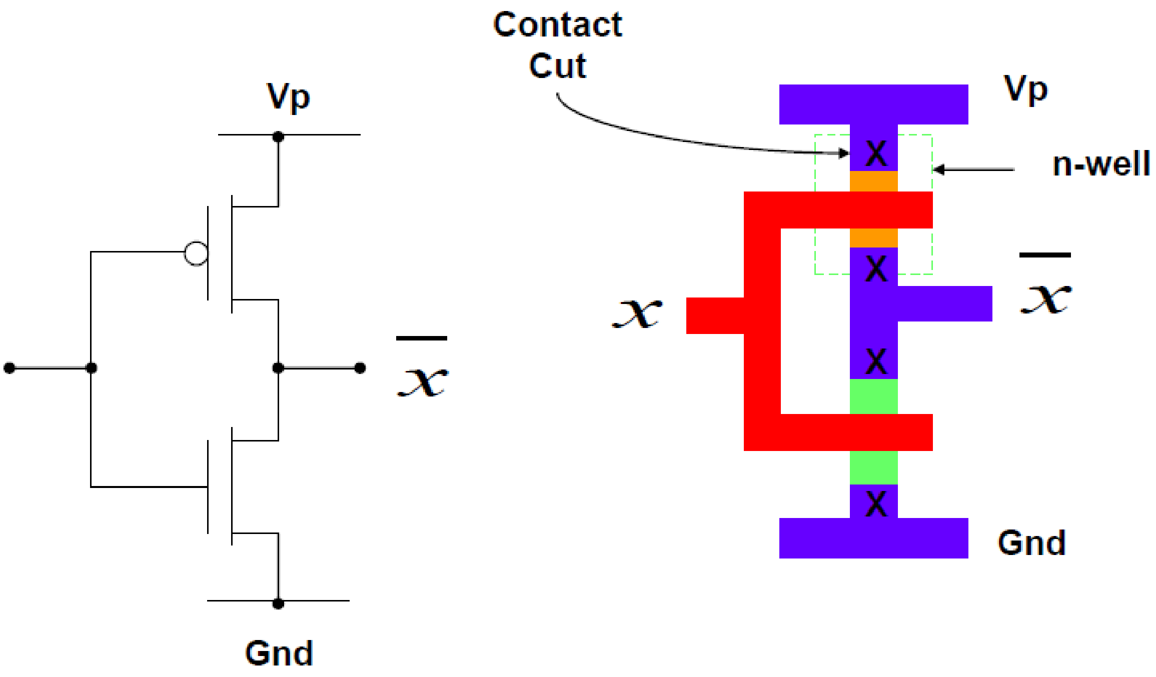


Both the power supply and ground are routed using the Metal layer

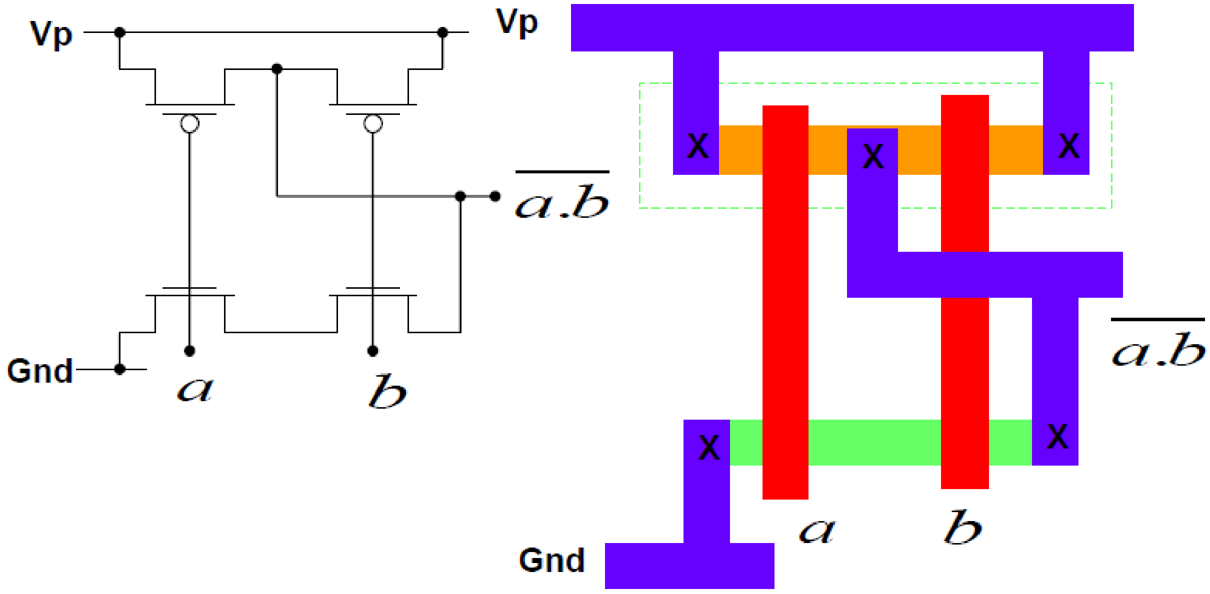
n+ and p+ regions are denoted using the same fill pattern. The only difference is the n-well

Contacts are needed from Metal to n+ or p+

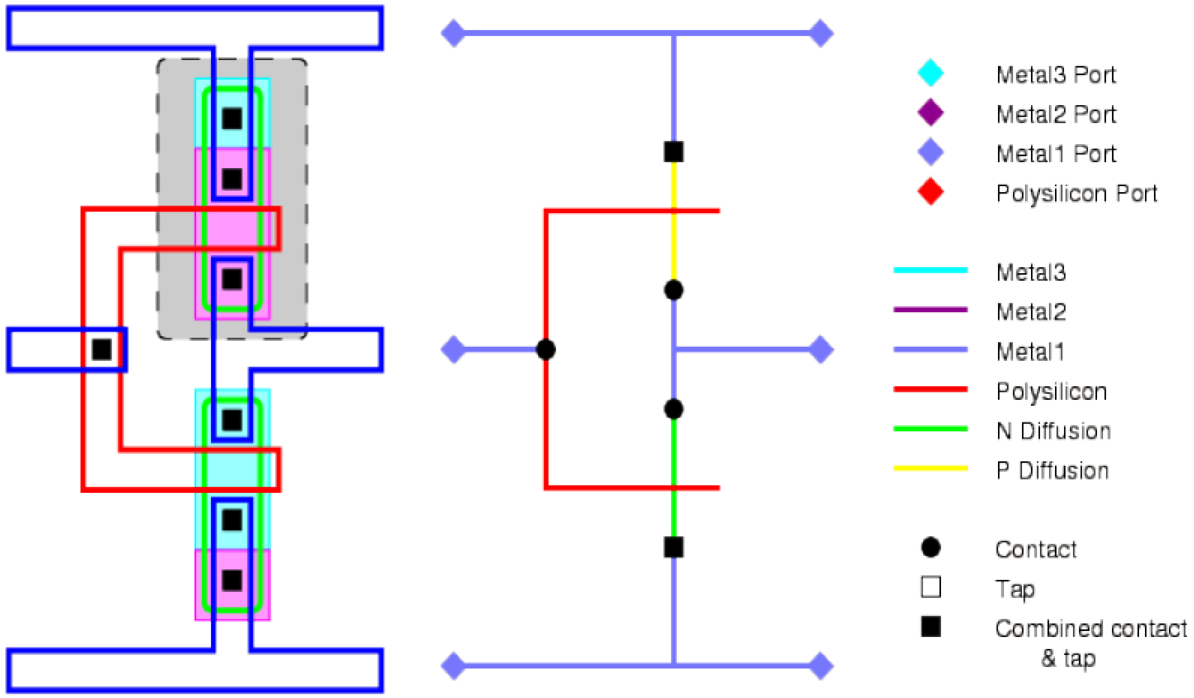
The CMOS NOT Gate



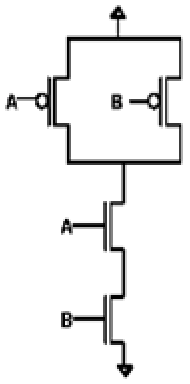
The CMOS NAND Gate

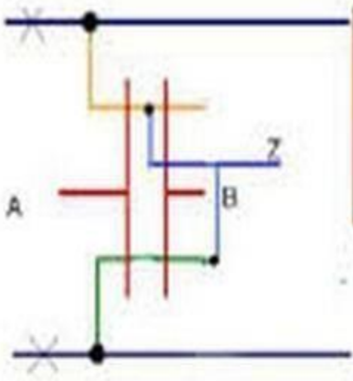


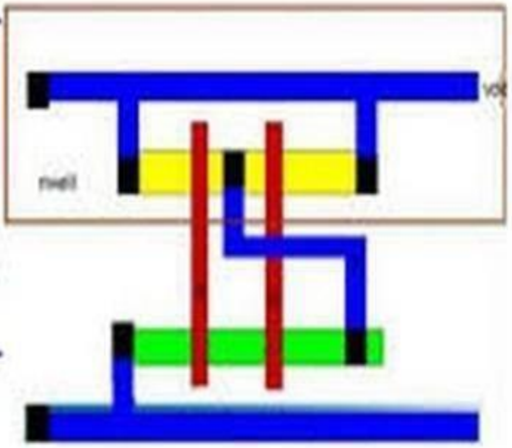
Layout&StickDiagramofCMOSInverter



2inputNAND gate

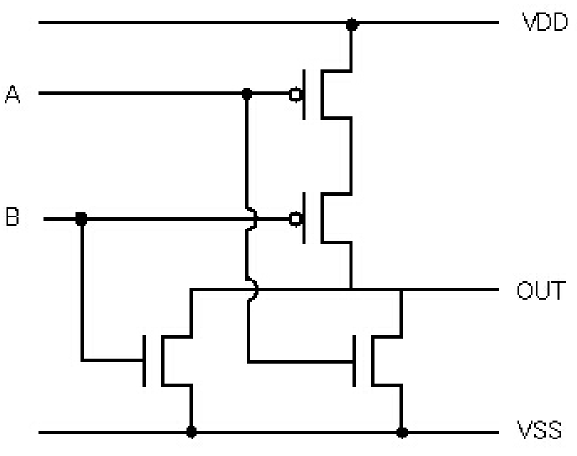




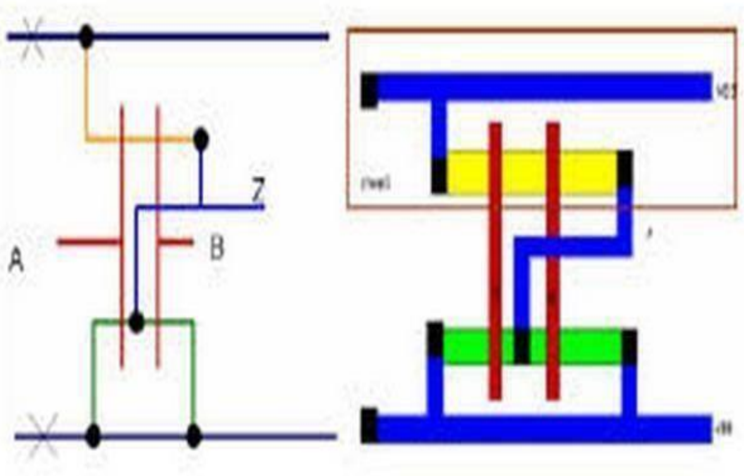


surface.

2inputNORgate



NOR gate in CMOS

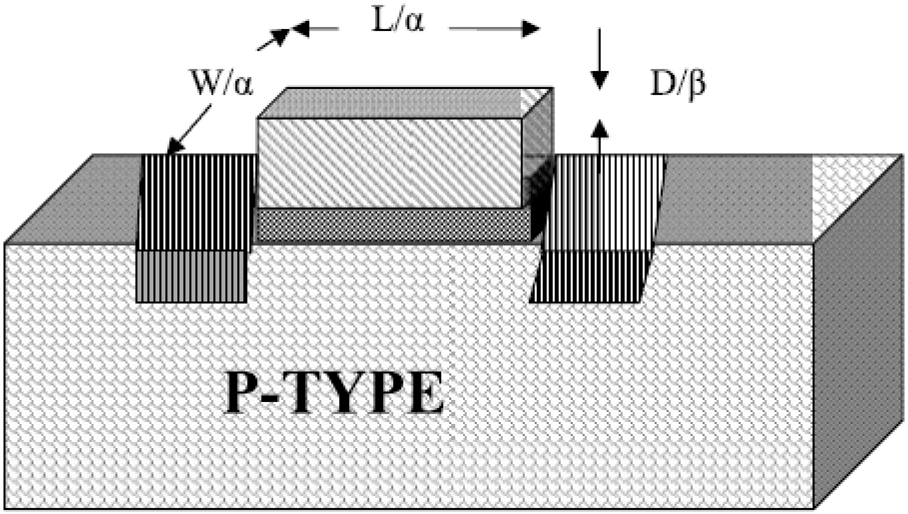


ScalingofMOScircuits

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

SCALING MODELS AND SCALING FACTORS:

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



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

SCALING FACTORS FOR DEVICE PARAMETERS:

GATE AREA AG :

$$A_g = L.W$$

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

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

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

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

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

GATE CAPACITANCE C_g:

$$C_g = C_O\ L.W$$

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

PARASITIC CAPACITANCE C_X:

C_X is proportional to A_X/d.

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

CARRIER DENSITY IN CHANNEL Q_{on}

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

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

Thus Q_{on} is scaled by 1.

CHANNEL RESISTANCE R_{on}

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

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

GATE DELAY T_d

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

Thus T_d is scaled by β^2/α^4

MAXIMUM OPERATING FREQUENCY F_O:

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

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

SATURATION CURRENT I_{DSS}:

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

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

CURRENTDENSITYJ:

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

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

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

SWITCHING ENERGY PER GATE E_g:

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

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

POWER DISSIPATION PER GATE P_g:

P_g comprise two components such that

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

Where the static component

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

And the dynamic component

$$P_{gd}=E_g f_o$$

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

POWER DISSIPATION PER UNIT AREA:

$$P_a = P_g / A_g$$

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

POWER-SPEED PRODUCT P_T:

$$P_T = P_g . T_d$$

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

LimitationsofScaling:

Scaling may cause a problem which prevents further miniaturization.

Substrate doping: -

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

(a) Substrate doping scaling factors:

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

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

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

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

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

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

Limits of miniaturization

The minimum size of transistor is determined by both process technology and the physics of the device itself.

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

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

Where

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

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

V=effective voltage across the junction

$V = V_{a+}V_B$

q=electron charge

N_B=doping level of substrate.

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

V_B=built-in (junction) potential

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

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

Depletion width

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

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

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

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

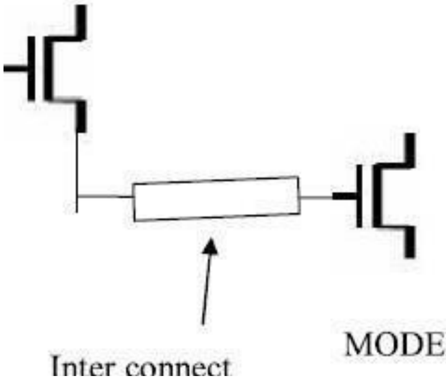
Limits due to interconnect and contact resistance

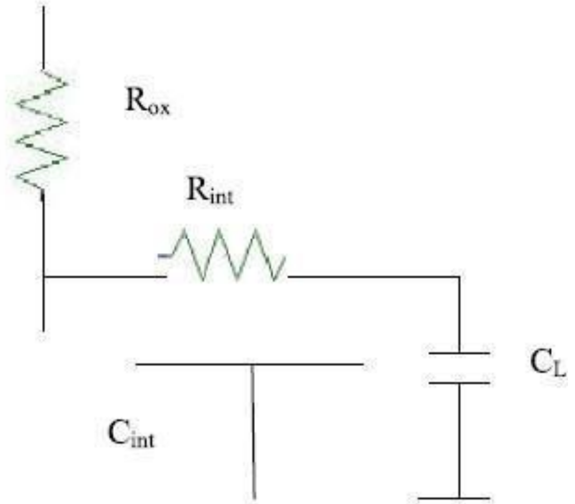
Since the width, thickness and spacing are scaled by $1/\alpha$, cross-section area must be scaled by $1/\alpha^2$ .Thus R is increased by α and I is scaled by $1/\alpha$. so IR drop remains constant. Thus driving capability and noise margins are degraded.

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

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

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





MODEL OF METAL INTERCONNECT

Now

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

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

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

R_{int} is the resistance of the interconnect

C_{int} is the capacitance of interconnect

t_{ox} is the thickness of dielectric oxide.

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

Assignmentquestions:

1. Drawthecircuitdiagram;stickdiagramandlayout forCMOSinverter.
2. Explainaboutthevariouslayoutdesignrules.
3. DrawthestaticCMOSlogiccircuitforthefollowingexpression
    1. i)Y=(ABCD)′ii)Y=[D(A+BC)]′
4. ExplainindetailaboutthescalingconceptinVLSIcircuitDesign.
5. DrawtheLayoutDiagramsforNANDGateusingnMOS..
6. Explainλ-basedDesignRulesinVLSIcircuitDesign.
7. DrawtheLayoutDiagramsforCMOSInverter.
8. Discussaboutthestickdiagramsandtheircorrespondingmasklayoutexamples
9. Drawthestickdiagramofp-wellCMOSinverterandexplaintheprocess.
10. Explainaboutthe2µmCMOSDesignrules anddiscusswithalayoutexample.
11. Drawandexplainthe layoutforCMOS2-inputNANDgate.
12. DrawtheflowchartofVLSIDesignflowand explaintheoperationofeachstepindetail.
13. DrawthestickdiagramforthreeinputAND gate.
14. Whatisthepurpose ofdesignrule?What is thepurpose ofstickdiagram?Whatarethe different approaches for describing the designrule? Give three approaches for making contacts between poly silicon and discussion in NMOS circuit.

UNITIII

GATELEVELDESIGNANDBASICCIRCUIT CONCEPTS

GatelevelDesign:

- Logicgatesandothercomplexgates
- Switchlogic
- Alternategatecircuits

BasicCircuitConcepts:

- SheetResistanceRsanditsconceptsto MOS
- AreaCapacitancescalculations
- InverterDelays
- Fan-inandfan-out.

CMOSLogicgatesandothercomplexgates

Name	Logic symbol	Logic equation
INVERTER		Out=~in;
AND		Out=a&b;
NAND		Out=~(a.b);
OR		Out=(a|b);
NOR		Out=~(a|b);
XOR		Out=a^b;
XNOR		Out=~(a^b);

CMOSlogicgateconcept:

The structure of a CMOS logic gate is based on complementary networks of n-channel and p-channel MOS circuits. Recall that the pMOS switch is good at passing logic signal '1', while nMOS switches are good at passing logic signal '0'. The operation of the gate has two main configurations:

- the nMOS switch network is closed, the output s=0 (figure 6-6 left)
- the pMOS switch network is closed, the output s=1 (figure 6-6 right)

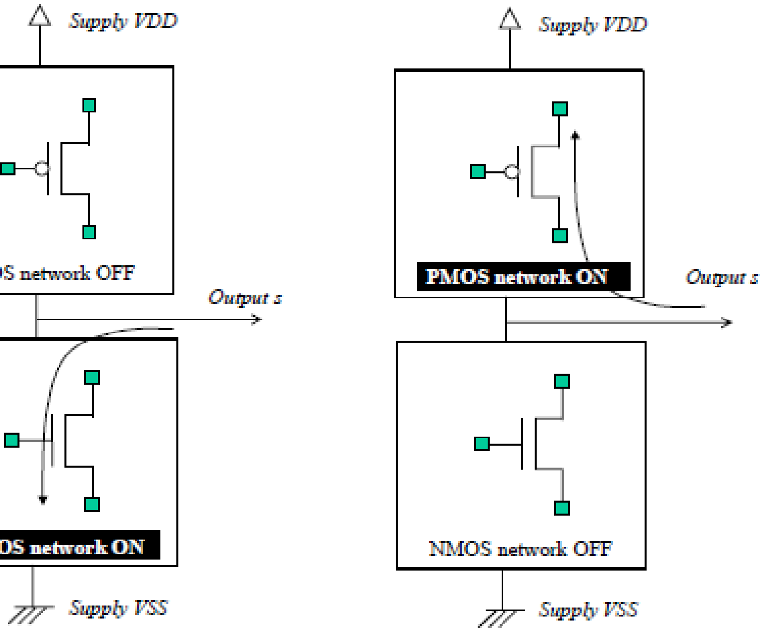


Fig. 6-6. General structure of a CMOS basic gate

Using complementary pairs of nMOS and pMOS devices, either the lower nMOS network is active, which ties the output to ground, either the upper pMOS network is active, which ties the output to VDD. In conventional CMOS basic gates, there should exist no combination when both nMOS and pMOS networks would be ON. If this case had happened, a resistive path would be created between VDD and VSS supply rails. The situation where neither nMOS and pMOS networks would be OFF should also be avoided, because the output would be undetermined.

CMOSStaticlogic

Static, fully complementary CMOS gate designs using inverter, NAND and NOR gates can build more complex functions. These CMOS gates have good noise margins and low static power dissipation at the cost of more transistors when compared with other CMOS logic designs. CMOS static complementary gates have two transistor nets (nMOS and pMOS) whose topologies are related. The pMOS transistor net is connected between the power supply and ie logic gate output, whereas the nMOS transistor topology is connected between the output id ground (Fig. 5.1). We saw this organisation with the NAND and NOR gates, but we point ut this topology to lead to a general technique to convert Boolean algebra statements to MOS electronic circuits.

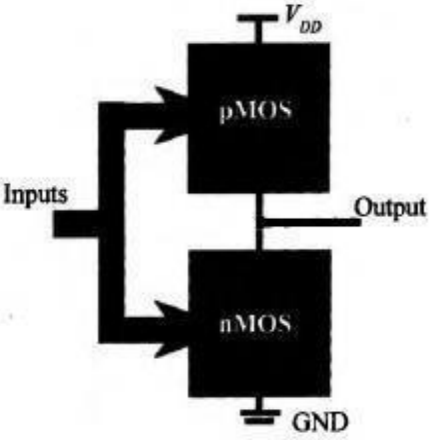
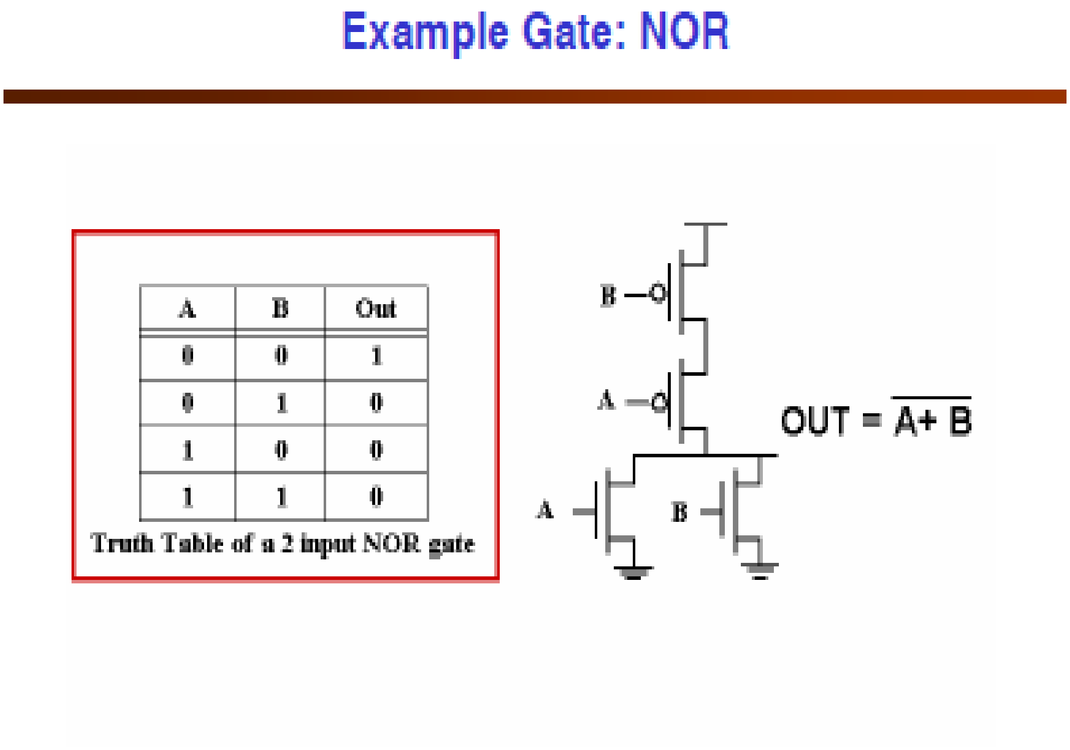


Fig. 5.1 Standard configuration of a CMOS complementary gate.

DesignProcedure:

1. Derive the nMOS transistor topology with the following rules:
    - Product terms in the Boolean function are implemented with series-connected nMOS transistors.
    - Sum terms are mapped to nMOS transistors connected in parallel.
2. The pMOS transistor network has a dual or complementary topology with respect to the nMOS net. This means that serial transistors in the nMOS net convert to parallel transistors in the pMOS net, and parallel connections within the nMOS block are translated to serial connections in the pMOS block.
3. Add an inverter to the output to complete the function if needed. Some functions are inherently negated, such as NAND and NOR gates, and do not need an inverter at the output state. An inverter added to a NAND or NOR function produces the AND and OR function. The examples below require an inverter to fulfil the function.

Examples:

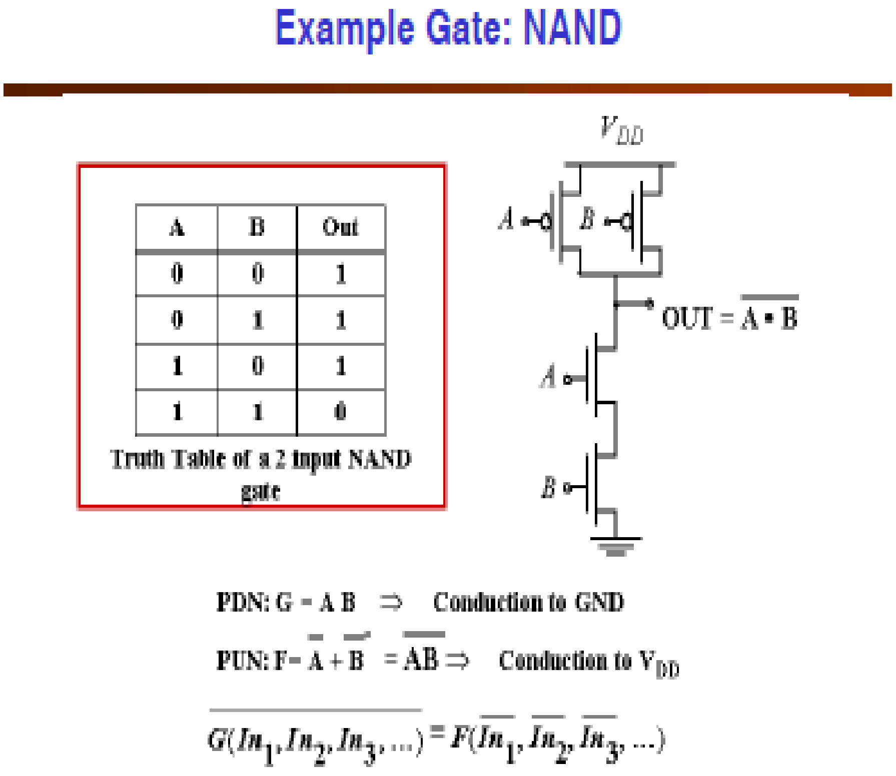


Example Gate: NOR

A	B	Out
0	0	1
0	1	0
1	0	0
1	1	0

Truth Table of a 2 input NOR gate

OUT = $\overline{A+B}$



Example Gate: NAND

A	B	Out
0	0	1
0	1	1
1	0	1
1	1	0

Truth Table of a 2 input NAND gate

OUT = $\overline{A \bullet B}$

PDN: G = A B ⇒ Conduction to GND

PUN: F = $\bar{A} + \bar{B}$ = $\overline{AB}$ ⇒ Conduction to V_{DD}

$$\overline{G(In_1, In_2, In_3, \ldots)} = F(\overline{In_1}, \overline{In_2}, \overline{In_3}, \ldots)$$

1.

Design a complementary static CMOS XOR gate at the transistor level. The XOR gate Boolean expression F has four literals and is $F = \bar{x}y + x\bar{y}$.

F is the sum of two product terms. The design steps are:

1. Derive the nMOS transistor topology with four transistors, one per literal in the Boolean expression. The transistors driven by $\bar{x}$ and y are connected in series, as well as the devices driven by x and $\bar{y}$. These transistor groups are connected in parallel, since they are additive in the Boolean function. The signals and their complements are generated using inverters (not shown). The nMOS transistor net is shown in Fig. 5.2.

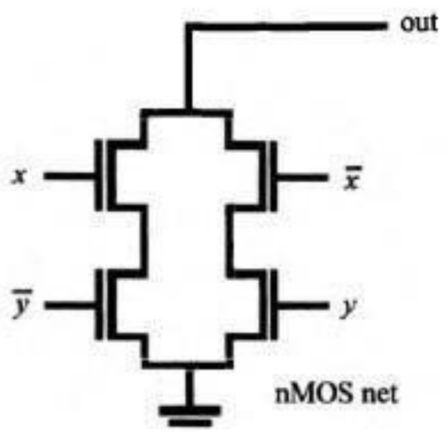


Fig. 5.2

2. Implement the pMOS net as a dual topology to the nMOS net. The pMOS transistors driven by $\bar{x}$ and y are connected in parallel, as are the devices driven by x and $\bar{y}$ (Fig. 5.3). These transistor groups are connected in series, since they are parallel connected in the nMOS net. The *out* node now implements $\bar{F}$.

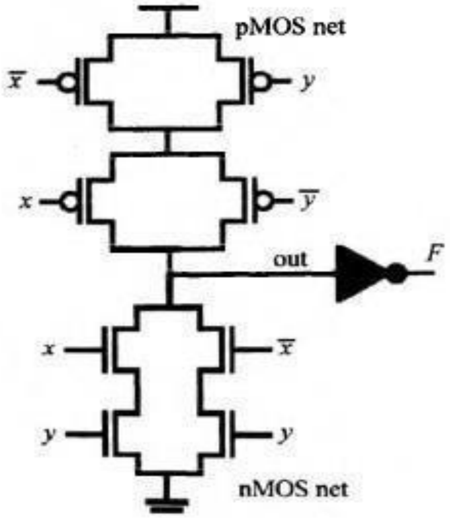


Fig. 5.3

3. Finally, add an inverter to obtain the function F, so that $F = \overline{\text{out}}$.

2. Design the nMOS transistor net for a Boolean function $F = x + \{\bar{y} \cdot [z + (t \cdot \bar{w})]\}$. We design this gate with a top-down approach. The nMOS transistor network is connected between the output and ground terminals, i.e., the lower box in Fig. 5.4(b). The higher-level function F is a sum of two terms:

$F = x +$ {operation A} where operation A stands for the logic within the brackets of F. The transistor version of this sum is shown in Fig. 5.4(a).

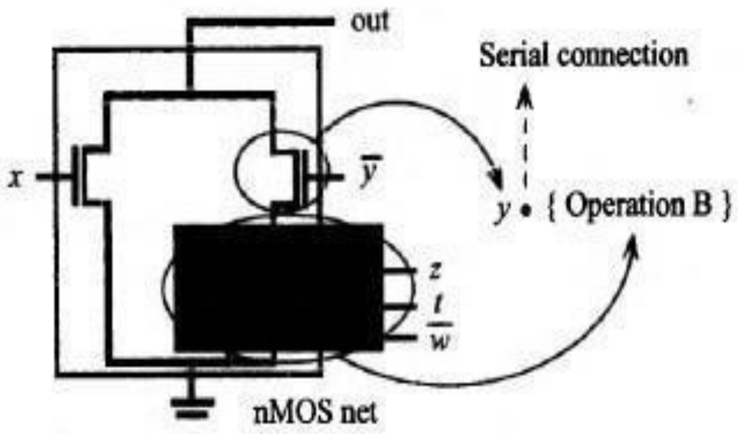


Fig. 5.4(a)

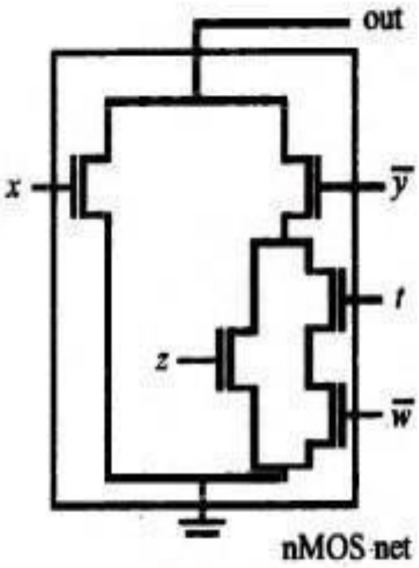


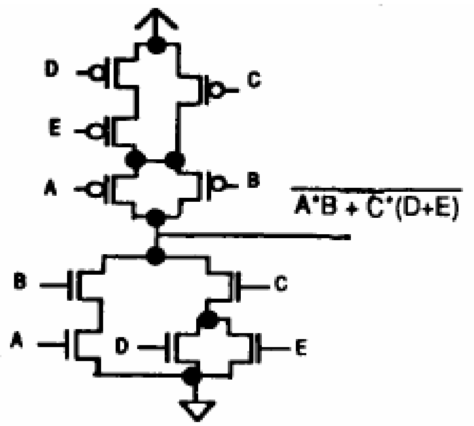
Fig. 5.4(b)

Hence, the design topology is a transistor controlled by input $\bar{y}$ in series with a third box that will implement *operation B*, as shown in Fig. 5.4. We then design the topology of box B. This is a transistor controlled by input z, in parallel with two transistors connected in series; one controlled by input t, and the other by input $\bar{w}$. The complete nMOS network is shown in Fig. 5.4(b). Once the nMOS block is designed, we build the pMOS block with a dual topological structure and then connect an inverter to its output, as shown in Fig. 5.6.

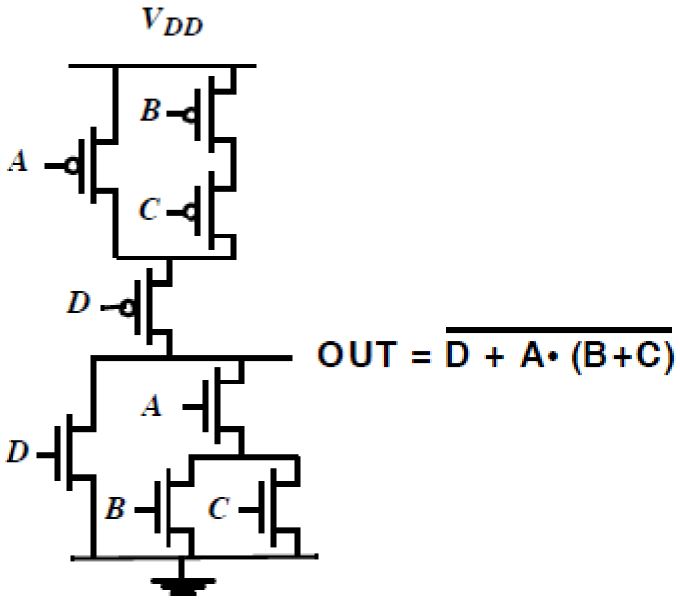
ComplexGates:

Complex Gate

- We can form complex combinational circuit function in a complementary tree. The procedure to construct a complementary tree is as follow:-
  - Express the boolean expression in an inverted form
  - For the n-transistor tree, working from the inner-most bracket to the outer-most term, connect the **OR** term transistors in parallel, and the **AND** term transistors in series
  - For the p-transistor tree, working from the inner-most bracket to the outer-most term, connect the **OR** term transistors in series, and the **AND** term transistors in parallel



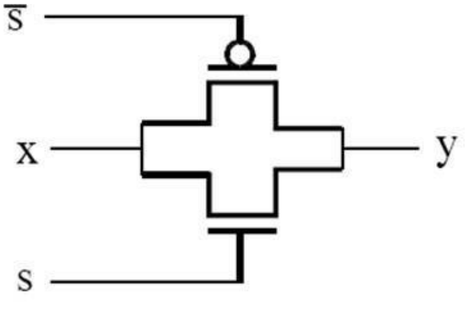
Example Gate: COMPLEX CMOS GATE



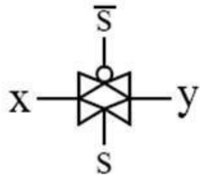
Transmissiongatelogic:

A **transmission gate** is an electronic element. It is a good non-mechanical relay, built with CMOS technology. Sometimes known as an analog gate, analog switch or electronic relay depending on its use. It is made by the parallel combination of an nMOS and a pMOS transistor with the input at the gate of one transistor being complementary to the input at the gate of the other.

A *transmission gate* is a essentiallya switchthat connects two points. Inorderto pass 0’s and 1’s equally well, a pair of transistors (one N-Channel and one P-Channel) is used as shown below:



Circuit



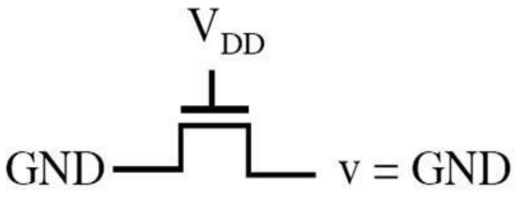
Symbol

Whens=1thetwotransistorsconductandconnect xandy

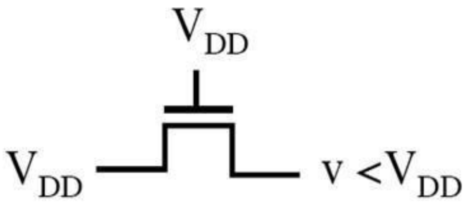
Thetoptransistorpassesxwhenit is1andthe bottomtransistor passesxwhenit is0 When s = 0

the two transistor are cut offdisconnecting x and y

N-ChannelMOSTransistorspassa0betterthana 1

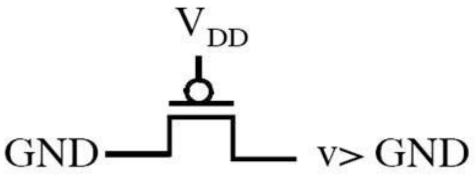


Passing 0

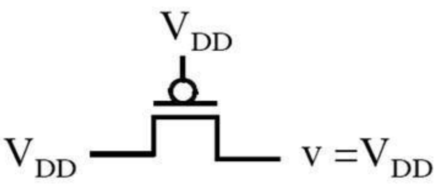


Passing 1

P-ChannelMOS Transistors passa1betterthana 0



Passing 0



Passing 1

This isthereasonthat N-Channeltransistorsareused inthepull-downnetwork and P-Channel in the pull-up network of a CMOS gate. Otherwise the noise margin wouldbesignificantly reduced.

Tristategates:

Many logic gates require a tri-state output—high, low, and high-impedance states. The high-impedance state is also called the high-Z state, and is useful when connecting many gate outputs to a single line, such as a data bus or address line. A potential conflict would exist if more than one gate output tried to simultaneously control the bus line. A controllable high-impedance-state circuit solves this problem.

There are two ways to provide high impedance to CMOS gates. One way provides tristate output to a CMOS gate by connecting a transmission gate at its output (Fig. 5.7). The control signal C sets the transmission gate conducting state that passes the non-tristated inverter output $\overline{\text{out}}$ to the tri-stated gate output *out*. When the transmission gate is off ($C = 0$), then its gate output is in the high-impedance or floating state. When $C = 1$, the transmission gate is on and the output is driven by the inverter.

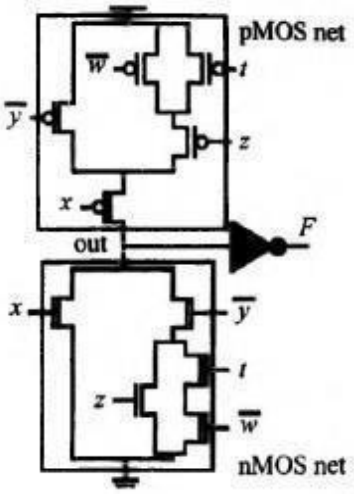


Fig. 5.6

A transmission gate connected to the output provides tri-state capability, but also consumes unnecessary power. The design of Fig. 5.7 contributes to dynamic power each time that the input and output ($\overline{\text{out}}$) are switched, even when the gate is disabled in the tri-state mode. Parasitic capacitors are charged and discharged. Since the logic activity at the input does not contribute to the logic result while the output is in tri-state, the power consumption related to this switching is **wasted.**

PassTransistorLogic

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 transistors.

Advantages are the low number of transistors and the reduction in associated interconnects. The drawbacks are the limited driving capability of these gates and the decreasing signal strength when cascading gates. These gates do not restore levels since their outputs are driven from the inputs, and not from V_{DD} or ground.

A typical CMOS design is the gate-level multiplexer (MUX) shown in Fig. 5.9 for a 2-to-1 **MUX**

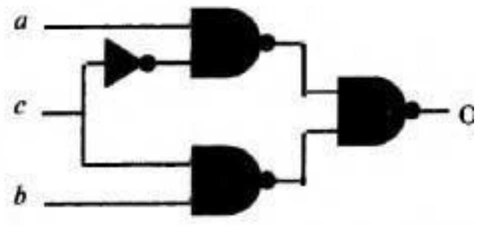
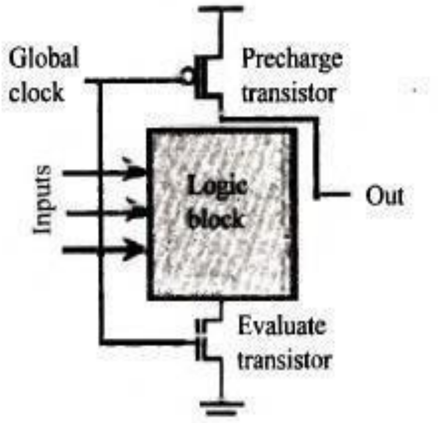


Fig. 5.9 (a) Standard 2-to-1 MUX design. (

DynamicCMOSlogic:



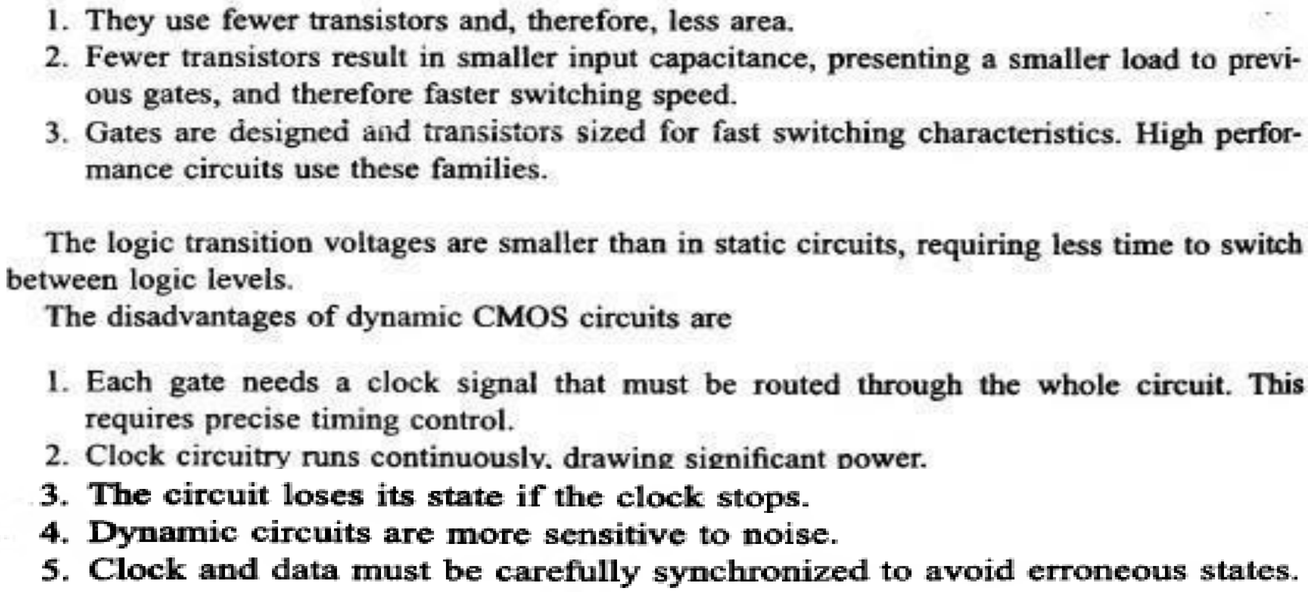
BasicStructure ofa dynamicCMOSgate

This logic looks into enhancing the speed ofthe pull up device by precharging the output node to V_{dd}. Hence we need to split the working of the device into precharge and evaluate stage for whichwe need a clock. Hence it is called as dynamic logic. The output node is precharged to V_{dd} by the pmos and is discharged conditionally through the nmos. Alternatively you can also have a p block and precharge the n transistor to V_{ss}. When the clock is low the precharge phase occurs. The pathto Vss is closed bythe nmos i.e. the ground switch. The pullup time is improved because of the active pmos which is already precharged. But the pull down time increases because of the ground switch.

Thereareafewproblemsassociated withthedesign,like

- Inputshavetochangeduringtheprechargestageandmustbestableduringthe evaluate. If this condition cannot occur then charge redistribution corrupts the output node.
- A simple single dynamic logic cannot be cascaded. During the evaluate phase the first gate willconditionallydischarge but bythe time the second gate evaluates, there is going to be a finite delay. Bythen the first gate may precharge.

MeritsandDemerits:



1. They use fewer transistors and, therefore, less area.
2. Fewer transistors result in smaller input capacitance, presenting a smaller load to previous gates, and therefore faster switching speed.
3. Gates are designed and transistors sized for fast switching characteristics. High performance circuits use these families.

The logic transition voltages are smaller than in static circuits, requiring less time to switch between logic levels.

The disadvantages of dynamic CMOS circuits are

1. Each gate needs a clock signal that must be routed through the whole circuit. This requires precise timing control.
2. Clock circuitry runs continuously, drawing significant power.
3. The circuit loses its state if the clock stops.
4. Dynamic circuits are more sensitive to noise.
5. Clock and data must be carefully synchronized to avoid erroneous states.

DominoCMOSLogic

This logic is the most common form of dynamic gates, achieving a 20%-50% performance increase over static logic. When the nMOS logic block discharges the out node during evaluation (Fig. 5.12), the inverter output out goes high, turning off the feedback pMOS. When out is evaluated high (high impedance in the dynamic gate), then the inverter output goes low, turning onthe feedbackpMOS device and providing a low impedance pathtoV_{DD}, This prevents the out node from floating, making it less sensitive to node voltage drift, noise and current leakage.

Domino CMOS allows logic gate cascading since all inputs are set to zero duringprecharge, avoiding erroneous evaluation from different delays. This logic allows static operation from the feedback latching pMOS, but logic evaluation still needs two sub cycles:precharge andevaluation.

Dominologicusesonlynon-invertinggates,making itanincompletelog family.Toachieveinverted logic, a separate inverting pathrunning in parallelwith the non inverted one must be designed.

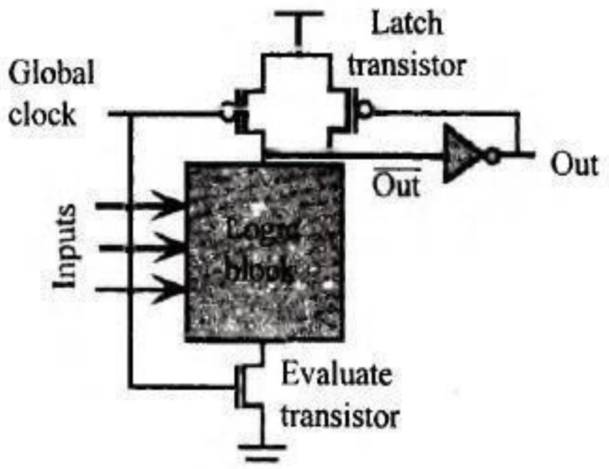


Fig. 5.12 Domino CMOS logic gate with feedback transistor.

Multiple output domino logic (MODL) is an extension of domino logic, taking internal nodes of the logic block as signal outputs, thus saving area, power, and performance. Compound domino logic is another design that limits the length of the evaluation logic to prevent charge sharing, and adds other complex gates as buffer elements (NAND, NOR, etc. instead of inverters) to obtain more area compaction. Self-resetting domino logic (SRCMOS) has each gate detect its own operating clock, thus reducing clock overhead and providing high performance.

NORA CMOS Logic. This design alternative to domino CMOS logic eliminates the output buffer without causing race problems between clock and data that arise when cascading dynamic gates. NORA CMOS (No-Race CMOS) avoids these race problems by cascading alternate nMOS and pMOS blocks for logic evaluation. The cost is routing two complemented clock signals. The cascaded NORA gate structure is shown in Fig. 5.13. When the global clock (*GC*) is low ($\overline{GC}$ high), the nMOS logic block output nodes are precharged high, while outputs of gates with pMOS logic blocks are precharged low. When the clock changes, gates are in the evaluate state.

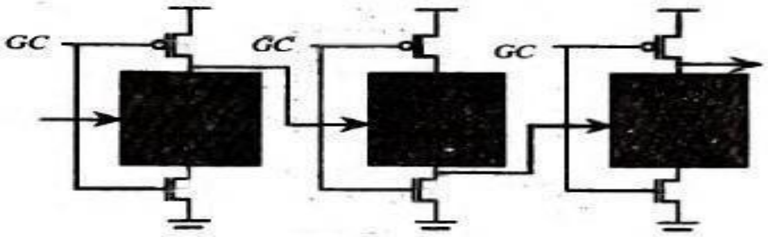
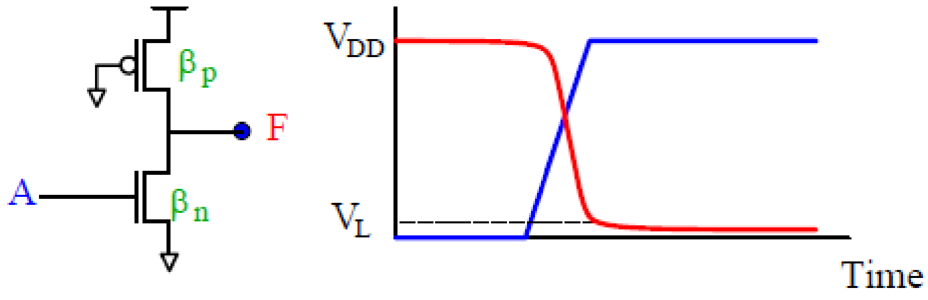


Fig. 5.13 NORA CMOS cascaded gates.

Pseudo–NMOSLogic:

pseudo-NMOS inverter



The inverter that uses a p-device pull-up or loads thathas its gate permanently ground. Ann-devicepull-downor driver isdrivenwiththe input signal. Thisroughlyequivalent to useof adepletionload is**NMOS**technologyand isthuscalled '**Pseudo-NMOS**'.Thecircuit isused ina variety of CMOS **logic** circuits.

The low output voltage can be calculated as

$$\beta_n(V_{DD}-V_{tn})V_L=\frac{\beta_P}{2}(V_{DD}-|V_{tp}|)^2$$

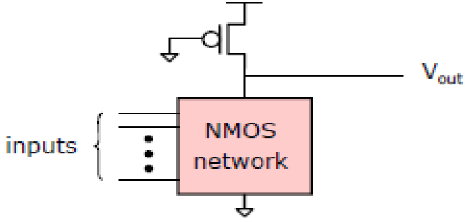
for $V_{tn}=-V_{tp}=V_t$

$$V_L=\frac{\beta_P}{2\beta_n}(V_{DD}-V_T)$$

Thus V_L depends strongly on the ratio β_p/β_n

The logic is also called ratioed logic

An N-input pseudo-NMOS gate



Features of pseudo-NMOS logic

- Advantages
  - ☐ Low area cost→only N+1 transistors are needed for an N-input gate
  - ☐ Low input gate-load capacitance→C_{gn}
- Disadvantage
  - ☐ Non-zero static power dissipation

BasicCircuitConcepts:

SheetResistanceRsanditsconceptstoMOS

Thesheet resistance isa measureofresistanceofthin filmsthat haveauniformthickness. It is commonly used to characterize materials made by semiconductor doping, metal deposition, resistive paste printing, and glass coating.

Example ofthese processes are: doped semiconductor regions (eg: siliconor polysilicon) and resistors.

Sheet resistance is applicable to two-dimensional systems where the thin film is considered to be a two-dimensional entity. It is analogous to resistivity as used in three-dimensional systems. When the term sheet resistance is used, the current must be flowing along the plane of the sheet, not perpendicular to it.

Model:

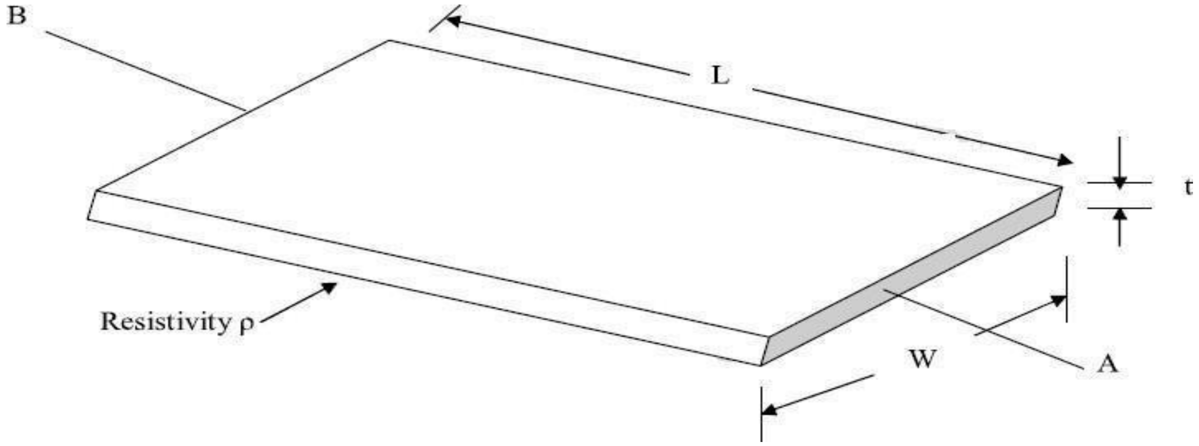
Consider a uniform slab of conducting material of resistivity ρ, of width W, thickness t, and length between faces L as shown below:

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

Where A = cross section area.

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

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



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

Where R_s =ohm per square or sheet resistance.

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

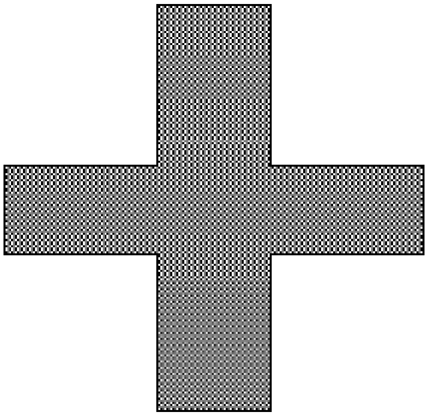
It is completely independent of the area of the square.

Typicalsheetresistance RsofMOSlayers

Layer	R_s ohm per square		
	5µm	Orbit	1.2µm
Metal	0.03	0.04	0.04
Diffusion	$10 \rightarrow 50$	$20 \rightarrow 45$	$20 \rightarrow 45$
Silicide	$2 \rightarrow 4$	-	-
Polysilicon	$15 \rightarrow 100$	$15 \rightarrow 30$	$15 \rightarrow 30$
n-transistor channel	10^4	2×10^4	2×10^4
p-transistor channel	2.5×10^4	4.5×10^4	4.5×10^4

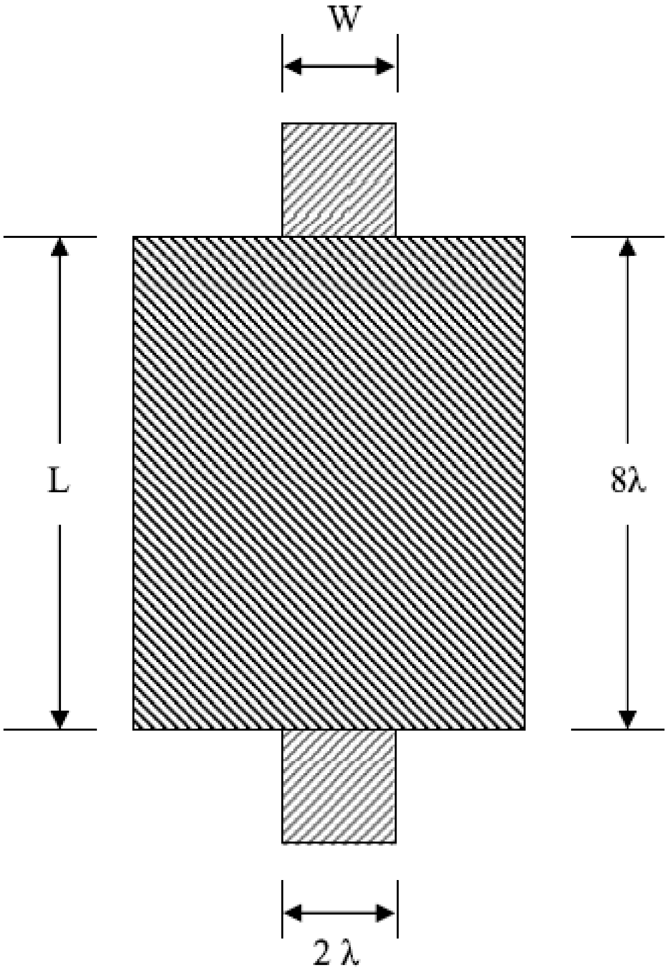
SHEET RESISTANCE CONCEPT APPLIED TO MOS TRANSISTORS AND INVERTERS

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



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

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



L = 8 λ and W = 2 λ

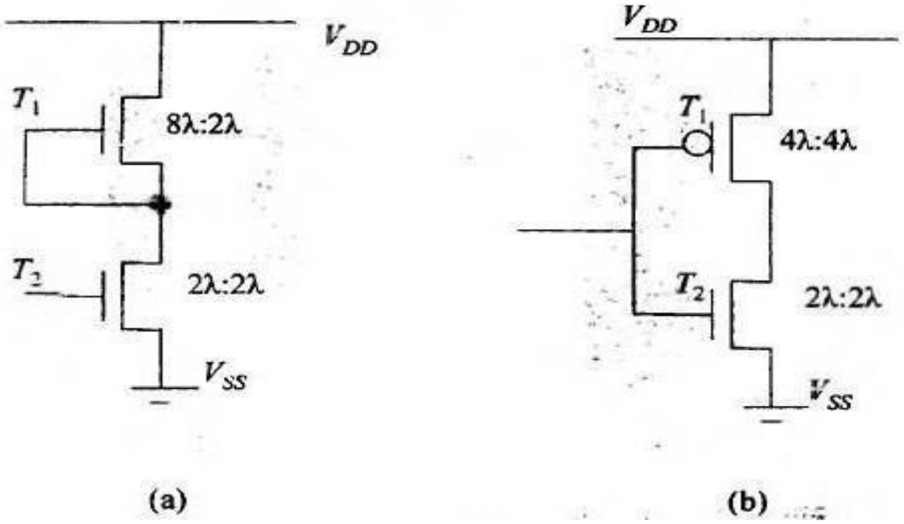
$$Z = \frac{L}{W} = 4$$

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

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

Calculation of ON Resistance of a Simple Inverter

Consider the simple nMOS inverter in Fig.



(a) (b)

Fig.(a)NMOSInverter(b) CMOSInverterresistancecalculations

- For the pull-up transistor (depletion mode MOSFET) the *L:W* value is 4:1, hence the value of *Z* is 4. R_{on} = 4 and value of on resistance is $4R_s$, i.e., 4×10^4 = 40 kΩ.
- Similarly, for the pull down transistor (enhancement mode MOSFET) the *L:W* value is 1:1 hence the value of *Z* is 1. R_{on} = 1 and value of resistance is $1R_s$, i.e., 1×10^4 = 10 kΩ.
- $Z_{p.u}$ to $Z_{p.d}$ = 4:1 hence the ON resistance between V_{DD} and V_{SS} is the total series resistance, i.e., 40 kΩ + 10 kΩ = 50 kΩ.

Consider the simple CMOS inverter in Fig.

- For the pull-up transistor (p-enhancement mode MOSFET) the *L:W* value is 1:1, hence, the value of *Z* is 4. R_{on} = 4 and value of on resistance is 4 R_s, i.e., $1 \times 25 \times 10^4$ = 25 kΩ (from the table value of R_s for p-channel transistor is 2.5×10^4 ohm/square).
- Similarly, for the pull down transistor (n-enhancement mode MOSFET) the *L:W* value is 1:1 hence the value of *Z* is 1. R_{on} = 1 and value of resistance is 1 R_s, i.e., 1×10^4 = 10 kΩ.
- In this case, there is no static resistance between V_{DD} and V_{SS} since at any point of time only one transistor is ON, but not both.
- When V_{in} = 1, the ON Resistance is 10 kΩ, when V_{in} = 0 the ON Resistance is 25 kΩ.

AreaCapacitancescalculations

From the concept of the transistors, we studied, it is apparent that as gate is separated from the channel by gate oxide an insulating layer, it has capacitance. Similarly, different interconnects run on the chip and each layer is separated by silicon dioxide.

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

Where

D = Thickness of silicon dioxide

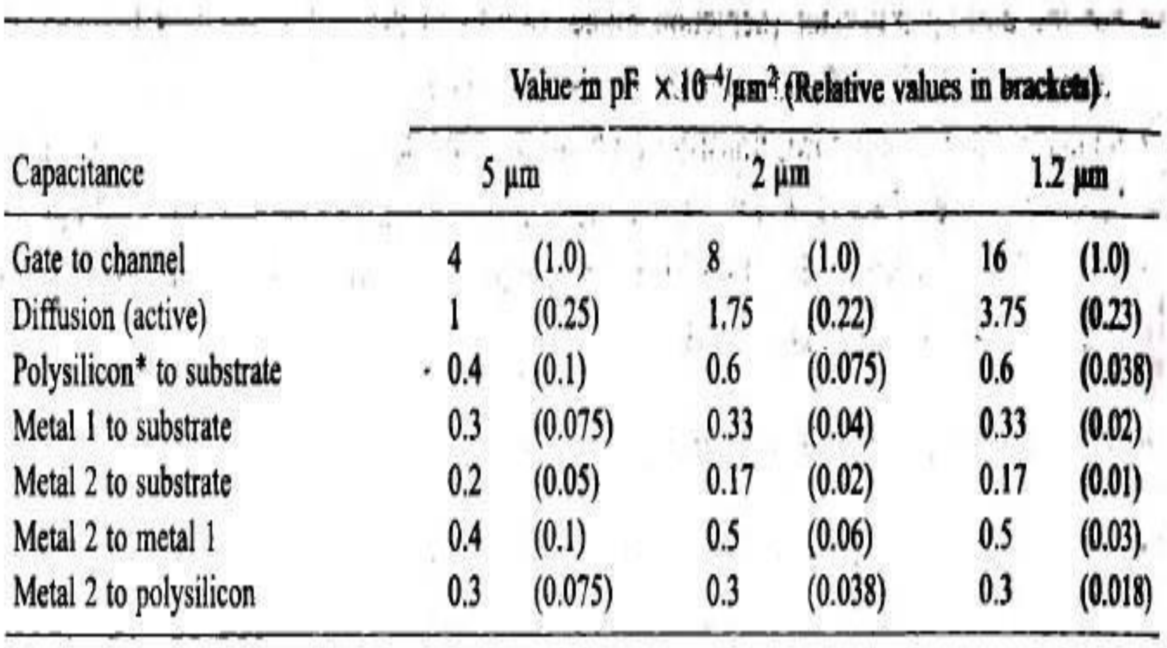
A = Area of plates

ε_{ins} = Relative permittivity of SiO_2 = 4.0

$\varepsilon_o = 8.85 \times 10^{-14}$ F/cm (permittivity of free space)

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

Typical values of area capacitance are given below in Fig. :



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

Standardunitofcapacitance:

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

C_g may be evaluated for any MOS process.

For example, for 5µm MOS circuits

Area/standard square = 5µm X 5µm = 25μm^2

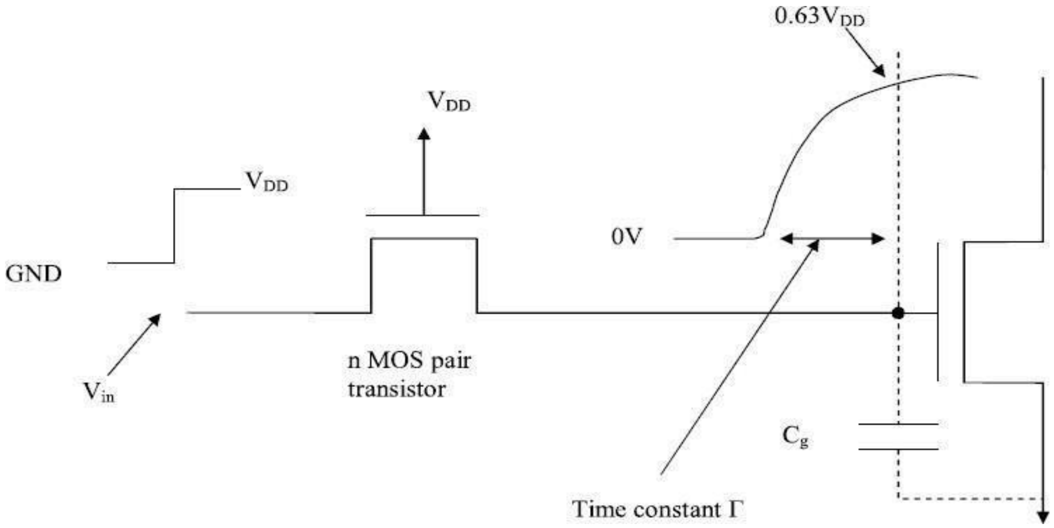
Capacitance value $= 4 \text{ X } 10^{-4}$ pF/μm^2
Thus standard value of $C_g = 25\ \mu m^2 \text{ X } 4 \text{ X } 10^{-4}$ pF/μm^2
$= 0.01$ pF

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

Calculation of Delay unit τ

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

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



For 5µm technology
$\Gamma = 10^4$ ohm X 0.01 pF
$= 0.1$ n sec

For 2µm technology
$\Gamma = 2 \text{ X } 10^4$ ohm X 0.0032 pF
$= 0.064$ n sec

For 1.2µm (orbit) technology
$\Gamma = 2 \text{ X } 10^4$ ohm X 0.0023 pF
$= 0.046$ n sec

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

$$\tau \approx \tau_{sd} = \frac{L^2}{\mu_n V_{ds}} = \frac{25\ \mu m^2 V \sec}{650\ cm^2 \quad 3V} X \frac{10^9\ n \sec\ cm^2}{10^8\ \mu m^2}$$
$$= 0.13 \text{ n sec}$$

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

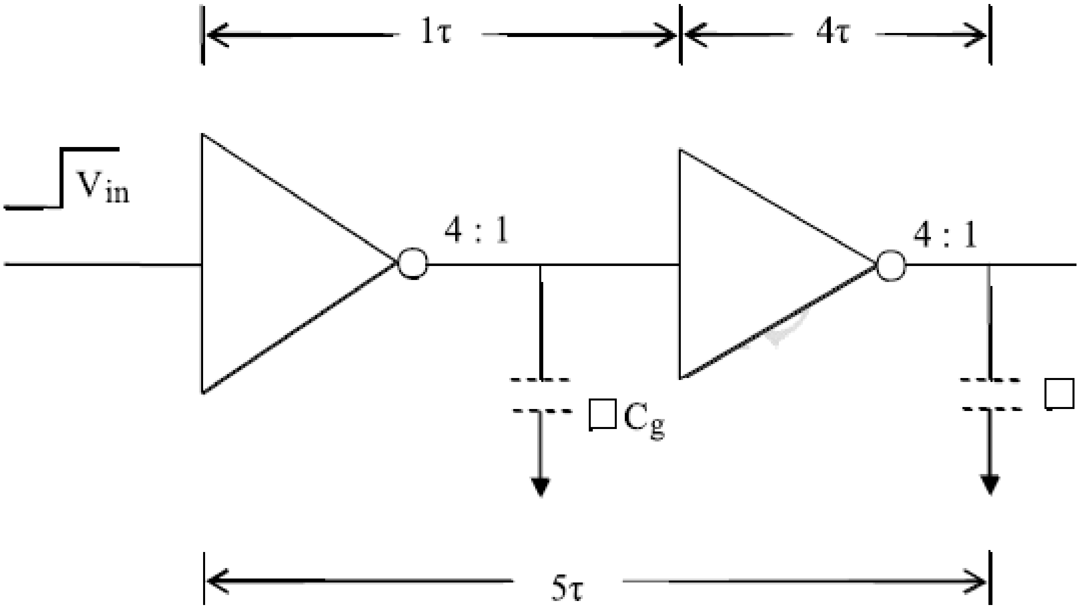
Inverter Delays:

nMOS Inverter Pair Delay

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

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

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



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

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

Consider CMOS inverter, the nmos rule does not apply. The gate capacitance is

$2\,C_g$ Because the input is connected to both transistor gates.

Minimum Size CMOS Inveter Pair Delay

When considering CMOS inverters, the nMOS ratio rule no longer applies, but we must allow for the natural (R_s) asymmetry of the usually equal size pull-up p-transistors and the n-type pull-down transistors. Figure 5.21 shows the theoretical delay associated with a pair of minimum size (both n- and p-transistors) lambda-based inverters. Note that the gate capacitance ($=2\square C_g$) is double that of the comparable nMOS inverter since the input to a CMOS inverter is connected to both transistor gates. Note also the allowance made for the differing channel resistances.

The asymmetry of resistance values can be eliminated by increasing the width of the p-device channel by a factor of two or three, but it should be noted that the gate input capacitance of the p-transistor is also increased by the same factor. This, to some extent, offsets the speed-up due to the drop in resistance, but there is a small net gain since the wiring capacitance will be the same.

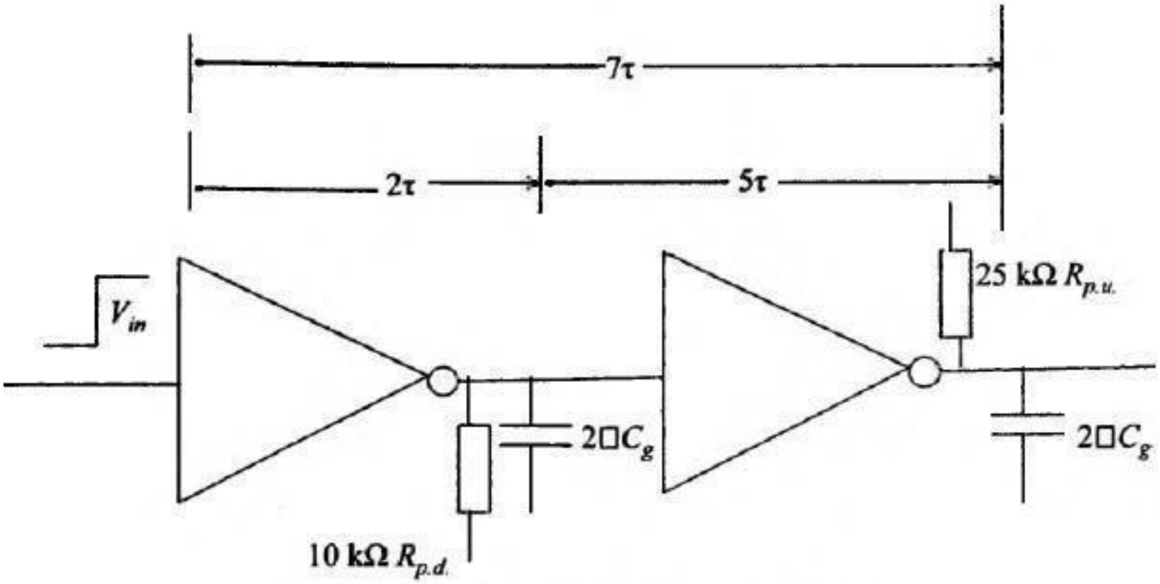


Fig. 5.21 Minimum size CMOS inverter pair delay.

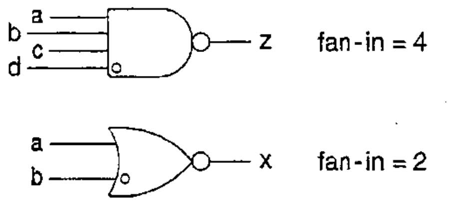
FaninandFanout:

- Fan-In=Number ofinputstoalogicgate
  - 4inputNANDhasaFI=4
  - 2inputNORhasaFI=2,etc.(SeeFig.abelow.)
- Fan-Out(FO)= Numberofgateinputswhicharedrivenbyaparticulargateoutput
  - FO=4 inFig.bbelowshowsanoutput wirefeedinganinputonfourdifferent logic gates
- Thecircuit delayofagateisa functionofboththeFan-Inand theFan-Out. Ex.m-

input NAND: $$\mathbf{t_{dr} = (R_p/n)(mC_d + C_r + kC_g)}$$

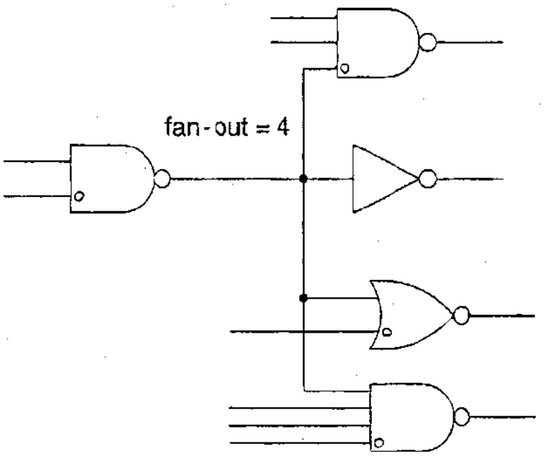
$$\mathbf{= t_{internal-r} + kt_{output-r}}$$

where n=widthmultiplier, m= fan-in, k= fan-out, Rp=resistanceofmin inverterPTx, Cg= gate capacitance, Cd = source/drain capacitance, Cr = routing (wiring) capacitance.



(a)

Note: The open circle adjacent to a logic gate input denotes the series transistor closest to the output.



(b)

•The circuit fall delay can be written in a similar manner.

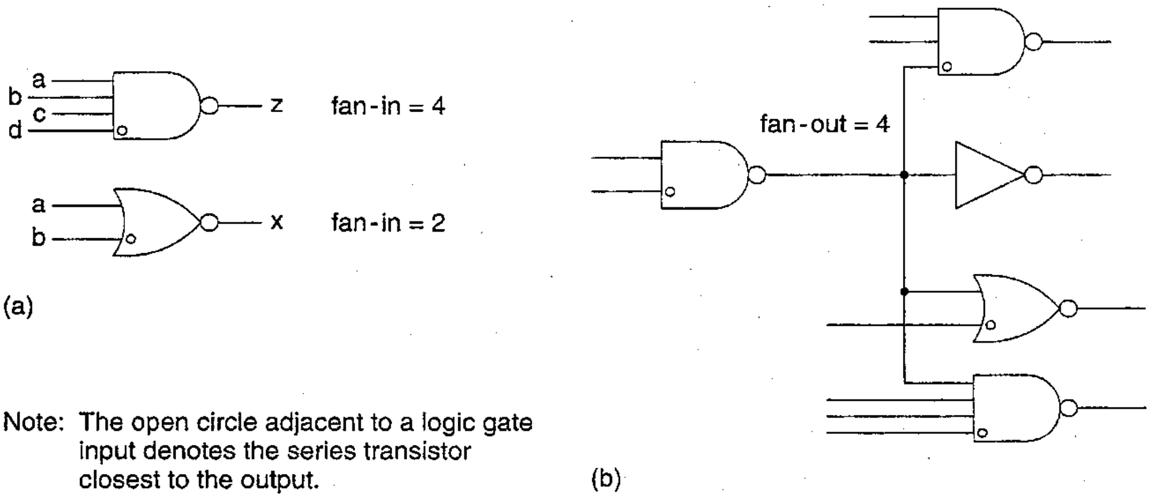
Ex. m-input NAND: $t_{df} = m(R_n/n)(mnC_d + C_r + kC_g)$

$$= t_{internal\text{-}f} + k\, t_{output\text{-}f}$$

where n = width multiplier, m = fan-in, k = fan-out, Rn = resistance of min inverter NMOS Tx, Cg = gate capacitance, Cd = source/drain capac, Cr = routing (wiring) capac.

If we set $t_{dr} = t_{df}$ for the case of symmetrical rise and fall delay, we obtain that Rp = m Rn and therefore,

$$\beta_p W_p = (\beta_n W_n)/m$$



(a)

Note: The open circle adjacent to a logic gate input denotes the series transistor closest to the output.

(b)

Summary

1. The sheet resistance is a measure of resistance of thin films that have a uniform thickness. It is commonly used to characterize materials made by semiconductor doping, metal deposition, resistive paste printing, and glass coating.
2. The resistance of the MOS layers depends on the thickness and the material of the layer. The resistance value of any square pattern is same as $W = L$.
3. Standard unit of capacitance is defined as gate to channel capacitance of a MOS transistor having $W = L$ = feature size that is standard.
4. Time constant $\tau = (1R_s$ (n channel) $\times 1\ \square C_g)$ seconds.

AssignmentQuestions:

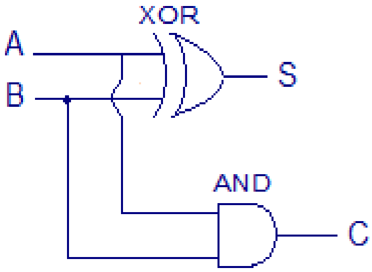
1. Describethe following:
    a) Pseudo-nMOSLogic
    b) DominoLogic.
2. Discussaboutthelogicsimplementedingateleveldesignandexplaintheswitchlogic implementation for a four way multiplexer.
3. Describeaboutthemethodsfordrivinglargecapacitiveloads.
4. Describeaboutthechoiceoffan–inandfan–outselectioningateleveldesign.
5. Whatarethealternategatecircuitsavailable?Explainanyoneofitemwithsuitable sketch by taking NAND gate as an example.
6. ExplaintheTransmissiongateand Tristateinverterlogic.
7. DescribethenMOSandCMOSinverterpairdelays.

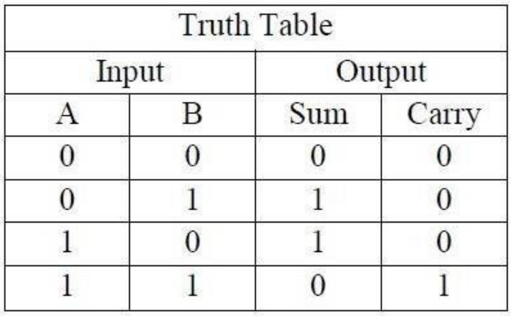
UNIT4

SubsystemDesign:

Datapath operators benefit from the structured design principles of hierarchy, regularity, modularity, and locality. They may use N identical circuits to process N-bit data. Related dataoperators are placed physically adjacentto each other toreducewirelength and delay. Generally, dataisarranged toflowin onedirection, whilecontrol signals are introduced in a direction orthogonal to the dataflow. Common data path operators include adders, one/zero detectors, comparators, counters, Boolean logic units, error-correcting code blocks, shifters, and multipliers.

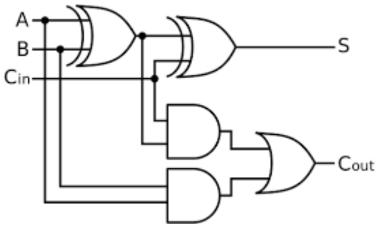
Adder: Addition forms the basis for many processing operations, from ALUs to address generation to multiplication to filtering. As a result, adder circuits that add two binary numbers are of great interest to digital system designers. Half adders and full adders forsingle-bit addition. The half adderadds two single-bit inputs, A and B. Theresultis 0, 1, or 2, sotwo bits arerequired torepresent thevalue; theyare called thesum S and carry-out Cout.

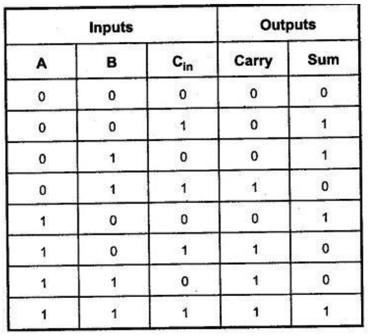




Truth Table			
Input		Output	
A	B	Sum	Carry
0	0	0	0
0	1	1	0
1	0	1	0
1	1	0	1

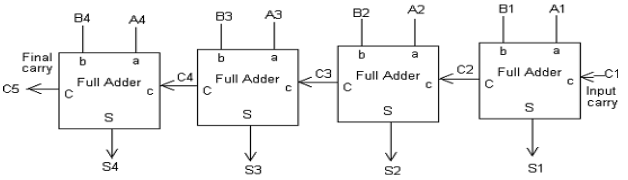
The carry-out is equivalent to a carry-in to the next more significant column of a multibit adder, so it can be described as having double the weight of the other bits. If multiple adders are to be cascaded, each must be able to receive the carry-in. Such a full adder hasa third input called C or Cin.





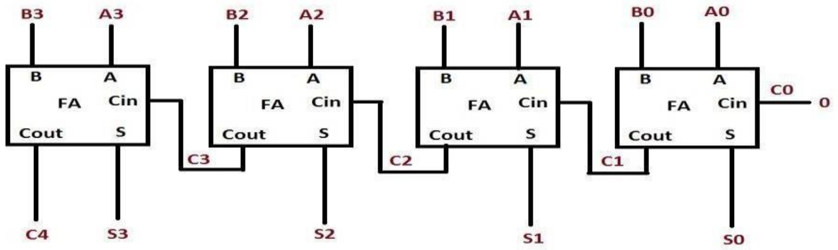
Inputs			Outputs	
A	B	C_{in}	Carry	Sum
0	0	0	0	0
0	0	1	0	1
0	1	0	0	1
0	1	1	1	0
1	0	0	0	1
1	0	1	1	0
1	1	0	1	0
1	1	1	1	1

N-bit adders take inputs {AN, ..., A1}, {BN, ..., B1}, and carry-in Cin, and compute the sum {SN, ..., S1} and the carry-out of the most significant bit Cout. They are calledcarry-propagate adders (CPAs)because the carry into each bit can influence the carryinto all subsequent bits.



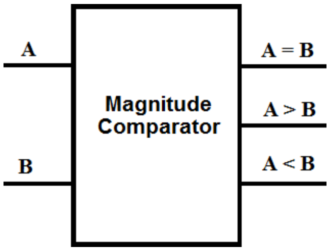
Carry-Ripple Adder: An N-bit adder can be constructed by cascading N full adders for N = 4. This is called acarry-ripple adder (or ripple-carry adder). The carry-out of bit i, Ci , is the carry-in to bit i + 1. This carry is said to have twice the weight of the sum Si. The delay of the adder is set bythe time for the carries toripple through the N stages, so the tCq Cout delayshould be minimized.

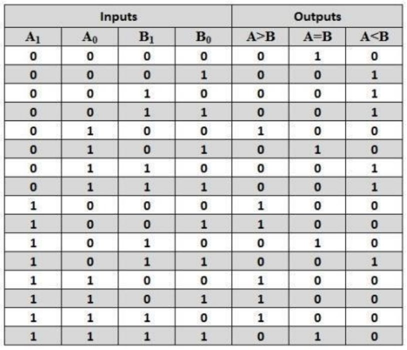
Every other stage operates on complementary data. The delay inverting the adder inputs or sum outputs is off the critical ripple-carrypath.



MagnitudeComparator

A magnitude comparator determines the larger of two binarynumbers. To compare two unsigned numbers A and B, compute B – A = B + A + 1. If there is a carry-out, A f B; otherwise, A > B. A zero detector indicates that the numbers are equal.





Inputs				Outputs		
A_1	A_0	B_1	B_0	A>B	A=B	A<B
0	0	0	0	0	1	0
0	0	0	1	0	0	1
0	0	1	0	0	0	1
0	0	1	1	0	0	1
0	1	0	0	1	0	0
0	1	0	1	0	1	0
0	1	1	0	0	0	1
0	1	1	1	0	0	1
1	0	0	0	1	0	0
1	0	0	1	1	0	0
1	0	1	0	0	1	0
1	0	1	1	0	0	1
1	1	0	0	1	0	0
1	1	0	1	1	0	0
1	1	1	0	1	0	0
1	1	1	1	0	1	0

Shifters:

Shifts can either be performed by a constant or variable amount. Constant shifts are trivial in hardware, requiring only wires. Theyarealso an efficient wayto perform multiplication or division bypowers of two. A variable shifter takes an N-bit input, A, a shift amount, k, and control signals indicating the shift type and direction. It produces an N-bit output, Y. There are three common types of variable shifts, each of which can be to the left or right: Rotate: Rotatenumbers in a circle such that emptyspots are filled with bits shifted offthe other end

○ Example:1011ROR1=1101;1011ROL1=0111

Logicalshift:Shiftthenumbertotheleftorrightandfillsemptyspotswithzeros.

○ Example:1011LSR1 =0101;1011LSL1 =0110

Arithmetic shift: Same aslogical shifter, but on right shifts fills themost significantbits withcopies of the sign bit (to properlysign, extend two's complementnumbers when using right shift byk for division by2k).

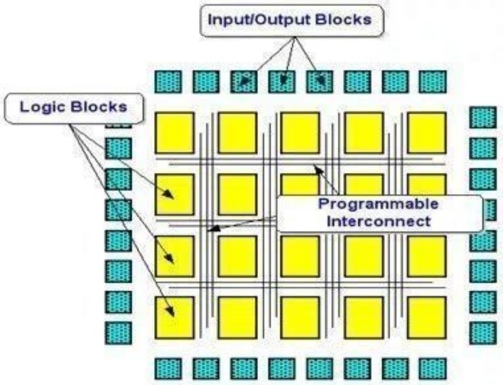
○ Example:1011ASR1= 1101;1011ASL1= 0110

Conceptually, rotation involves an array of N N-input multiplexers to select each of the outputs from each of the possibleinputpositions. This is called anarrayshifter. The arrayshifterrequires a decoderto producethe 1-of-Nhot shift amount. In practice, multiplexers with more than 4–8 inputs have excessive parasitic capacitance, so they are faster toconstruct from logvN levels of v-input multiplexers. Thisis called a logarithmic shifter.

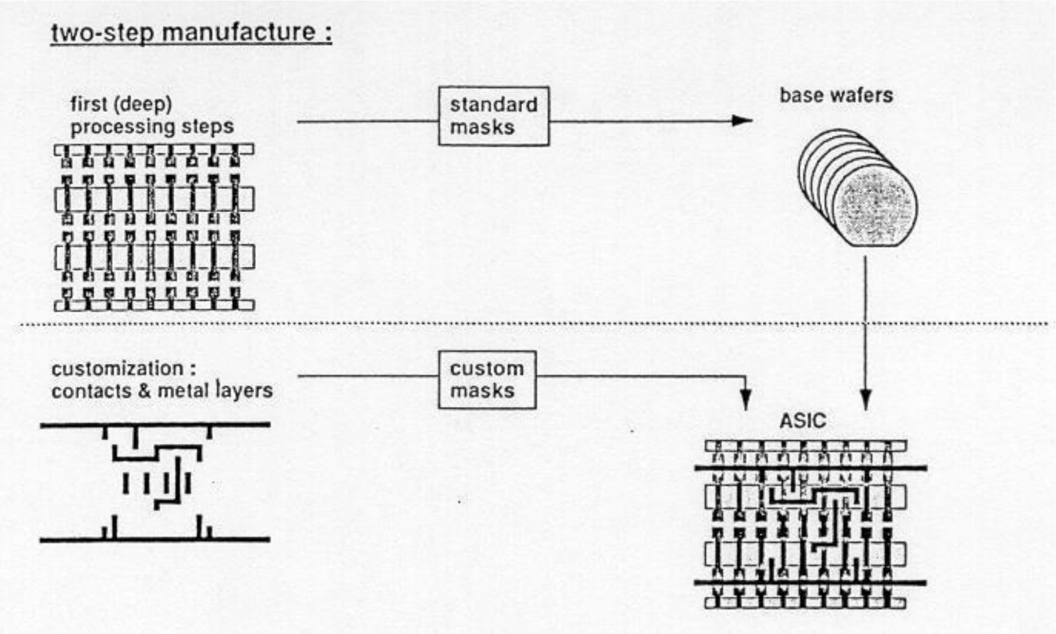
VLSIDesignStyles

Several design styles can be considered for chip implementation of specified algorithms or logic functions. Each design style has its own merits and shortcomings, and thus a proper choice has to be made by designers in order to provide the specified functionalityat low cost and in a timely manner.

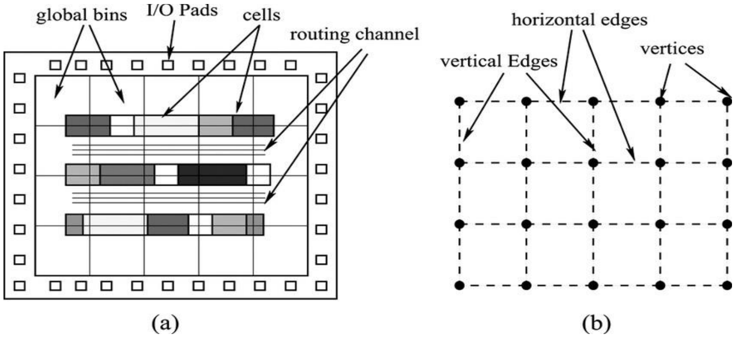
Field Programmable Gate Array (FPGA) Fully fabricated FPGA chips containing thousands or even more, of logic gates with programmable interconnects, are available to users for their custom hardware programming to realize desired functionality. This design style provides a means for fast prototyping and also for cost-effective chip design, especiallyfor low-volumeapplications.Atypical field programmable gatearray(FPGA) chipconsistsofI/O buffers, an arrayof configurable logic blocks (CLBs), and programmable interconnect structures. The programming of the interconnects is accomplished by programming of RAM cells whose output terminals are connected to the gates of MOS pass transistors. Thus, the signal routing between the CLBs and the I/O blocks is accomplished by setting the configurable switch matrices accordingly. The general architecture of an FPGA chip from Xilinx . showing the locations of switch matrices used for interconnect routing.



Gate Array Design :In termsof fast prototyping capability, the gatearray(GA) rankssecond after theFPGA with a typical turn-around time of a few days. While user programming is central to the design implementationof the FPGA chip, metal mask design and processing is used for GA. Gate array implementation requires a two-step manufacturing process: The first phase, which is based on generic (standard) masks, 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. Since the patterning of metallic interconnects is doneat theend ofthe chip fabricationprocess,theturn-around time canstillbe short, afewdays to a fewweeks. A corner of a gate arraychip which contains bonding pads on its left and bottom edges, diodes for 1O protection, nMOS transistorsand pMOS transistors forchip output drivercircuits adjacent to bonding pads, arrays of nMOS transistors and pMOS transistors, underpass wire segments, and power and ground buses along with contact windows. The availability of these routing channels simplifies the interconnections, even using one metal layer only. Interconnection patternsthat perform basic logic gates can be stored in a library, which can then be used to customize rows of uncommitted transistors according to the netlist.



Standard-Cells Based Design: The standard-cells based design is one of the most prevalent full custom design styles which require development of a full custom mask set. The standard cell is also called the polycell. In this design style, all of the commonlyused logic cells are developed,characterized, and stored in a standard cell library. A typical librarymaycontain a few hundred cells includinginverters, NAND gates, NOR gates, complex AOI, OAI gates, D latches, and flip-flops. Each gate type can be implemented in several versions to provide adequate driving capability for different fan-outs. For instance, the inverter gate can have standard size, double size, and quadruple sizesothatthechipdesigner can choosetheproper sizetoachievehigh circuit speedandlayout density. Each cell is characterized according to several different characterization categories, including: Delay time versus load capacitance, Circuit simulation model, Timing simulation model, Fault simulation model, Cell data for place-and-route, Maskdata. To enable automated placement of the cells androuting of inter-cell connections, eachcell layout is designed with a fixed height, so that a number of cells can be abutted side-by-side to form rows. The power and ground rails typically run parallel to the upper and lower boundaries of the cell, thus, neighboring cells share a common power andground bus. Theinputandoutputpinsare located on theupper andlower boundariesofthecell.



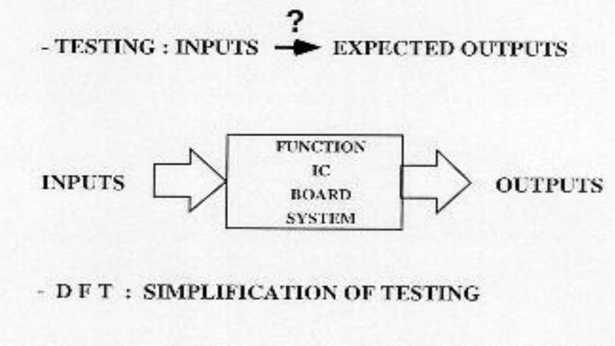
FullCustomDesign

Although the standard-cells based design style is sometimes called full custom design, in a strict sense, it is somewhat less thanfullycustomized since the cells are pre-designed forgeneral use and the same cells are utilized in manydifferent chipdesigns. In atrulyfull-custom design, theentiremask design is doneanew without useofany library. However, the development cost of such a design style is becoming prohibitively high. Thus, the concept of design reuse is becoming popular in order toreduce design cycle time and development cost. Themost rigorous full custom design can be the designof a memory cell, be it static ordynamic. Since the same layout design is replicated, there would not be any alternative to high density memory chip design. For logic chip design, a good compromise can be achieved by using a combination of different design styles on the same chip, such as standard cells, data-path cells and programmable logic arrays (PLAs). In real full-custom layout in which the geometry, orientation and placementofeverytransistor is doneindividuallybythedesigner, design productivityis usuallyvery low - typically a few tens of transistors per day, per designer. In digital CMOS VLSI, full-custom design is rarely used due to the high laborcost. Exceptions to this include the design of high-volume products such as memory chips, high-performance microprocessors and FPGA masters. Figure 14.23 shows the full layout of the Intel 486 microprocessor chip, which is a good example of a hybrid fullcustom design. Here, one can identify four different design styles on onechip:memorybanks(RAM cache), data-path unitsconsisting of bit-slice cells, control circuitry mainlyconsisting of standard cells and PLA blocks.

UNIT-V
CMOSTESTING

Needfortesting

Designof logic integratedcircuits inCMOS technologyis becoming moreandmorecomplex since VLSI is the interest of many electronic IC users and manufacturers. A common problem to be solved by both users and manufacturers is the testing of these ICs.



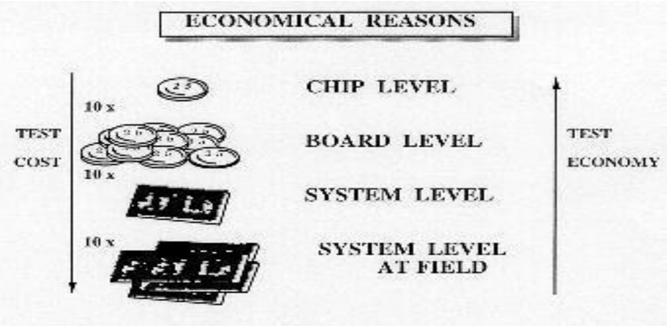
Testing can be expressed by checking if the outputs of a functional system (functional block, Integrated Circuit, Printed Circuit Board or a complete system) correspond to the inputs applied to it. If thetest of this functional system is positive, then thesystem is good for use. If the outputs are different than expected, then the system has a problem: so either the system is rejected (Go/No Go test), or a diagnosis is applied to it, in order to point out and probably eliminate the problem's causes.

Testing is applied to detect faults after several operations : design, manufacturing, packaging and especiallyduringtheactivelife of a system, andthus sincefailures causedbywear-out canoccur at any moment of its usage.

Design for Testability (DFT) is the ability of simplifying the test of any system. DFT could be synthesized by a set of techniques and design guidelines where the goals are :

- minimizingcostsofsystemproduction
- minimizingsystemtest complexity:test generationandapplication
- improvingquality
- Avoidingproblems oftimingdiscordanceorblocknatureincompatibility.

Intheproductionprocess cycle, a fault canoccur at the chip level. If a test strategyis consideredat the beginning of the design, then the fault could be detected rapidly, located and eliminated ata very low cost. When the faulty chip is soldered on a printed circuit board, the cost of fault remedy would be multiplied by ten. And this cost factors continues to apply until the system has been assembled and packaged and then sent to users.



ManufacturingTests:

Whereas verification or functionality tests seek to confirm the function of a chip as a whole, manufacturing tests are used to verify that every gate operates as expected. The need to do this arises from a number of manufacturing defects that might occur during either chip fabrication or acceleratedlifetesting(wherethechip is stressedby over-voltageandover-temperatureoperation). Typical defects include the following:

- Layer-to-layershorts(e.g.,metal-to-metal)
- Discontinuouswires(e.g.,metalthinswhencrossingverticaltopologyjumps)
- Missingor damagedvias
- Shorts through the thin gate oxide to the substrate or well

Theseinturnleadtoparticular circuitmaladies, includingthefollowing

- Nodesshortedtopower or ground
- Nodesshortedtoeachother
- Inputsfloating/outputsdisconnected

Tests are required to verify that each gate and register is operational and has not been compromised by a manufacturing defect. Tests can be carried out at the wafer level to cull out bad dies, or can be left until the parts are packaged. This decision would normally be determined by the yield and package cost. If the yield is high and the package cost low (i.e., a plastic package), then the part can be tested only once after packaging. However, if the wafer yield was lower and the package cost high (i.e., an expensive ceramic package), it is more economical to first screen bad dice at the wafer level. The length of the tests at the wafer level can be shortened to reduce test time based on experience with the test sequence.

Apart from the verification of internal gates, I/O integrity is also tested,with the following tests being completed:

- I/Olevels(i.e., checkingnoisemarginfor TTL,ECL, orCMOSI/Opads)
- Speedtest

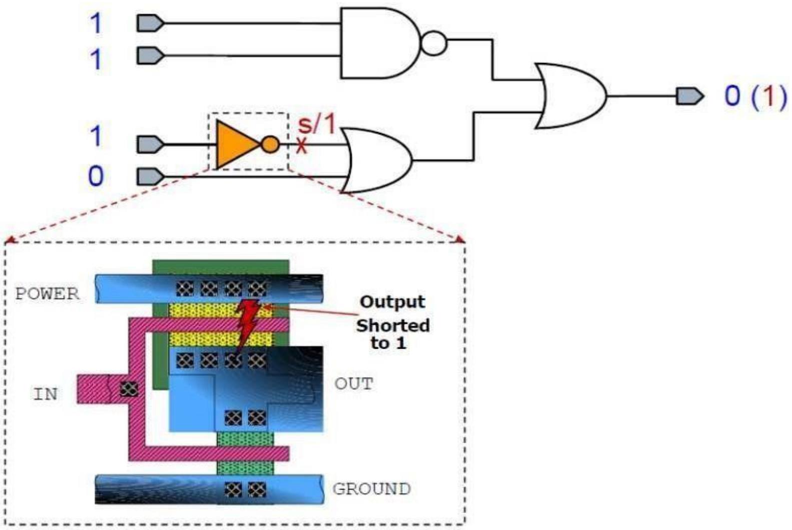
With the use of on-chip test structures described here, full-speed wafer testing can be completed witha minimumof connectedpins. This can beimportant inreducingthecost ofthe wafer test fixture.

In general, manufacturing test generation assumes the function of the circuit/chip is correct. It requires ways of exercising all gate inputs and monitoring all gate outputs.

Test Principles: Thepurposeof manufacturingtest is toscreen out most of the defectiveparts beforetheyareshippedtocustomers. Typicalcommercialproducts target a defect rateof350– 1000 defects per million (DPM) chips shipped.

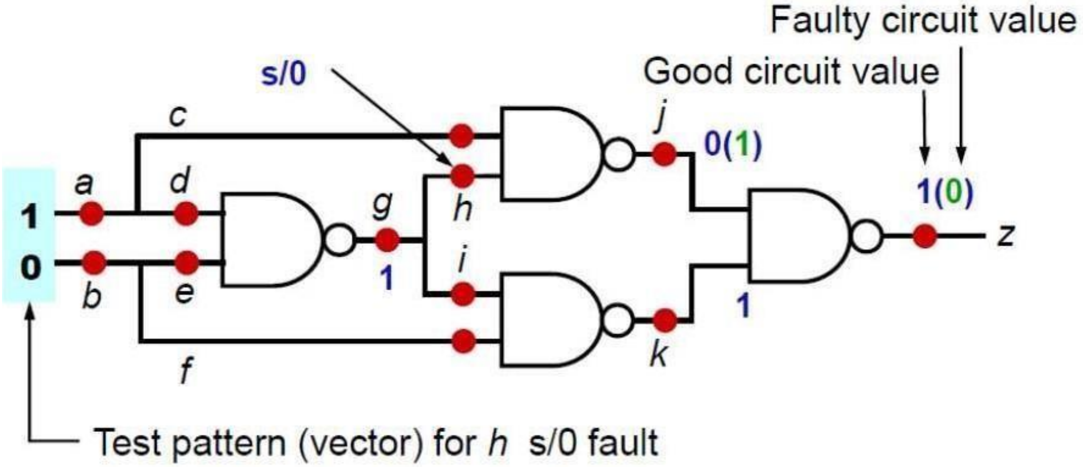
FaultModels

To deal with the existence ofgood and bad parts, it is necessary to propose a fault model; i.e., a model for how faults occur and theirimpact on circuits. The most popular model is called the Stuck-Atmodel. The Short Circuit/ Open Circuit model can be a closer fit to reality, but is harder to incorporate into logic simulation tools.



Stuck-At Faults In the Stuck-At model, a faulty gate input is modeled as a stuck at zero (Stuck-At-0, S-A- 0) or stuck at one (Stuck-At-l, S-A-l). This model dates from board-level designs,where itwasdetermined to be adequate formodeling faults.Figure illustrates how an S-A-0 or S-A-1 fault might occur. These faults most frequently occurdue to gate oxide shorts (the nMOS gate to GND or the pMOS gate to VDD) or metal-to-metal shorts.

Short-Circuit and Open-Circuit Faults Other models include stuck-open or shorted models. Two bridging or shorted faults are shown in Figure. The short S1 results in an S-A-0 fault at input A, while short S2 modifies the function of the gate.



It is evident that to ensure the most accurate modeling, faults should be modeled at the transistor level because it is only at this level that the complete circuit structure is known. For instance, in the case of a simple NAND gate, the intermediate node between the series nMOS transistors is hidden by the schematic. This implies that test generation should ideally take accountof possible shorts and open circuitsat the switch level.Expediencydictatesthatmost existingsystemsrelyonBooleanlogicrepresentationsofcircuitsandstuck-atfaultmodeling.

Observability

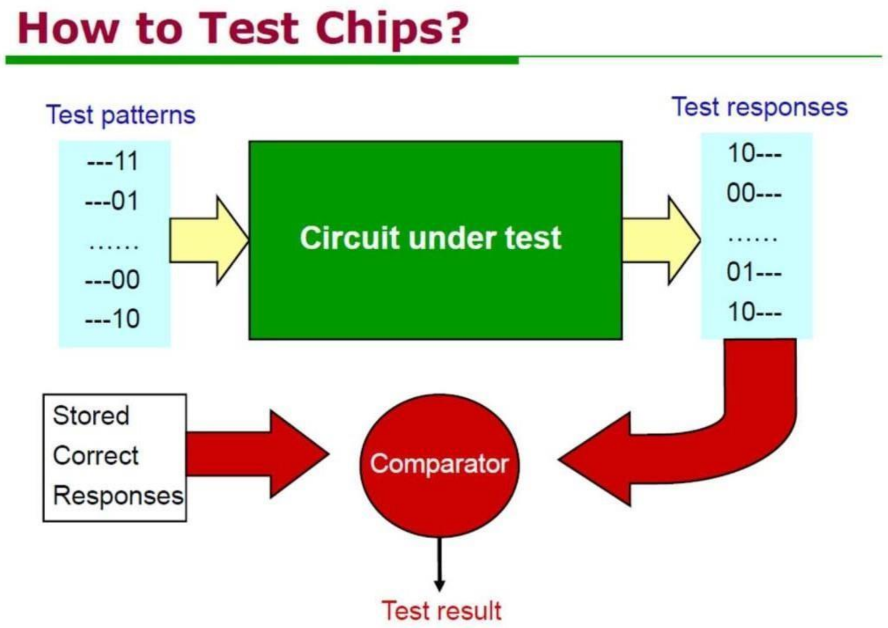
The observability of a particular circuit node is the degree towhich youcan observe that node at theoutputs of an integrated circuit (i.e., thepins). This metric is relevant when you want to measure the output of a gate within a larger circuit to check that it operates correctly. Given the limited number of nodes that can be directly observed, it is the aim of good chip designers to have easily observed gate outputs. Adoption of some basic design for test techniques can aid tremendously in this respect. Ideally, you should be able to observe directly or with moderate indirection (i.e., you may have to wait a few cycles) every gate output within an integrated circuit. While at one time this aim was hindered by the expense of extra test circuitry and alack ofdesignmethodology,currentprocessesanddesignpracticesallowyou to approach this ideal.

Controllability

The controllability of an internal circuit node within a chip is a measure of the ease of setting the node to a 1 or 0 state. This metric is of importance when assessing the degree of difficulty of testing a particular signal within a circuit. An easily controllable node would be directly settable via an input pad. Anode with little controllability, such as the most significant bit of a counter, might require many hundreds or thousands of cycles to get it to the right state. Often, you will find it impossible to generate a test sequence to set a number of poorly controllable nodes into the right state. It should be the aim of good chip designers to make all nodes easily controllable.In common with observability, the adoption of some simple design for test techniques can aid in this respect tremendously. Making all flip-flops resettable via a global reset signal is one step toward good controllability.

FaultCoverage

A measure of goodness of a set of test vectors is the amountof fault coverage it achieves. That is, for the vectors applied, what percentages of the chip's internal nodes werechecked? Conceptually, the way in which the fault coverage is calculated is as follows. Each circuitnodeistaken in sequence andheld to0(S-A-0),andthecircuitissimulated withthetest vectors comparing the chipoutputswith a known goodmachine—a circuit withno nodes artificially set to 0 (or1). When adiscrepancy isdetectedbetweenthefaulty machine and thegood machine, the fault is marked as detected and the simulation is stopped. This is repeatedforsetting the nodeto 1(S-A-1).In turn,everynode isstuck (artificially)at 1 and 0 sequentially. The fault coverage of a set of test vectors is the percentage of the total nodesthatcanbedetectedasfaulty whenthevectorsareapplied.Toachieveworld-class quality levels, circuits are required to have in excess of 98.5% fault coverage. The Verification MethodologyManual is the bible for fault coverage techniques.



IDDQ TESTING: It is a simple method to identify defects on IC on the steady state power supply current. It is also a method for testing CMOS integrated circuits for the presence of manufacturing faults. It relies on measuring the supply current (Idd) in the quiescent state (whenthe circuitisnotswitchingandinputsare heldatstaticvalues).The currentconsumed in the state is commonly called Iddq for Idd (quiescent) and hence the name. Iddq testing uses the principle that in a correctly operating quiescent CMOS digital circuit, there is no static current pathbetweenthepower supplyandground, except for a smallamount of leakage. Many common semiconductor manufacturing faults will cause the current to increase by orders of magnitude, which can be easily detected. This has the advantage of checking the chip for many possible faults with one measurement.Another advantage is that itmay catch faultsthatare not found by conventional stuck-at fault test vectors. Iddq testing is somewhat more complex than just measuring thesupply current. If a lineis shorted to Vdd, for example, it will still draw no extra current if the gate driving the signal is attempting to set it to '1'. However, a different vector set that attempts to set the signal to 0 will show a large increase in quiescent current, signaling a bad part. Typical Iddq test vector sets may have 20 or so vectors. Note that Iddqtest vectors require only controllability, and not observability. This is because the observability is through the shared power supply connection. Iddq testing has many advantages:

□ Itisasimpleanddirecttestthatcanidentify physicaldefects.
□ Theareaanddesigntimeoverheadareverylow.

□ Testgenerationisfast.

□ Testapplicationtimeisfastsincethevectorsetsaresmall.

□ Itcatchessomedefectsthatothertests,particularlystuck-atlogic tests,donot.

Drawback:Comparedtoscantesting, Iddqtestingis time consuming, andthen more expensive, since is achieved by currentmeasurements that take muchmore time than reading digital pins in mass production.

AUTOMATICTESTPATTERNGENERATION(ATPG)

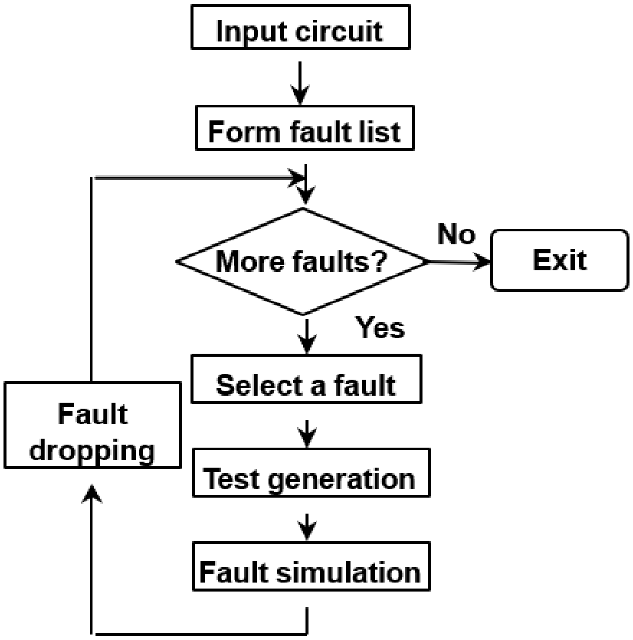
Historically,in the ICindustry,logicandcircuitdesignersimplementedthe functions at the RTL or schematic level, mask designers completed the layout, and test engineerswrote the tests. In many ways, the test engineers were the Sherlock Holmes of the industry, reverse engineering circuits and devising tests that would test the circuitsin an adequate manner. For the longest time, test engineers implored circuit designers to include extra circuitry to ease the burden of test generation. Happily, as processes have increased in density and chips have increased in complexity, the inclusion of test circuitry has become less of an overhead for both the designer and the manager worried about the cost of the die. In addition, as tools have improved, more of the burden for generating tests has fallen on the designer. To deal with this burden, Automatic Test Pattern Generation (ATPG) methods have been invented. The use of some form of ATPG is standard for most digital designs.

It is the process of generating test patterns for a given fault model. If we go by exhaustive testing, in the worst case, we may require 2n (where n stands for no. of primary inputs) assignments to be applied for finding test vector for a single stuck-at fault. It is impossible for us to manually use exhaustive testing or path sensitization method to generate a test pattern for chips consisting of millions of transistors. Hence,we needan automated process, a.k.a. Automatic Test Pattern Generation (ATPG).

A cycle of ATPG can generally be divided into two distinct phases: 1) creation of the test; and 2) application of the test. During the creation of the test, appropriate models for the device circuit are developed atgate or transistor levelin such away that the output responses of a faulty device for a given set of inputs will differ from those of a good device. This generationoftestisbasicallyamathematicalprocessthatcanbedoneinthreeways:

1)bymanualmethods;2)byalgorithmicmethods(withorwithoutheuristics);and
3) by pseudo- random methods. The software used for complex ATPG applications are quite expensive, but the process of generating a test needs to be done only once at the end of the design process.

When creating a test, the goal should be to make it as efficient inmemory space and time requirementsasmuch as possible.As such,the ATPG processmustgenerate the minimum or near minimum set of vectors needed to detect all the important faults of a device. The main considerations for test creation are: 1) the time needed to construct the minimal test set; 2) the size of the pattern generator, or hardware/software system needed to properly stimulate the devices under test; 3) the size of the testing process itself; 4) the time needed to load the test patterns; and 5) the external equipment required (if any).



DESIGNSTRATEGIESFOR TEST,

Design for Testability The keys to designing circuits that are testable are controllability and observability. Restated, controllability is the ability to set (to 1) and reset (to 0) every node internal to the circuit. Observability is the ability to observe, either directly or indirectly, the state of any node in the circuit. Good observability and controllability reduce the cost of manufacturing testing because they allow high fault coverage with relatively few test vectors. Moreover, they can be essential to silicon debug because physicallyprobinginternal signalshas become so difficult.

We will first cover three main approaches to what is commonly called Design for Testability (DFT). These may be categorized as follows:

Adhoctesting

Scan-based approaches

Built-inself-test(BIST)

AdHocTesting

Ad hoc test techniques, as their name suggests, are collections of ideas aimedat reducing the combinational explosion of testing. They are summarized here for historical reasons. They are only useful for small designs wherescan, ATPG, and BIST arenot available. A complete scan-based testing methodology is recommended forall digital circuits.Having said that, the following are common techniques for ad hoc testing:

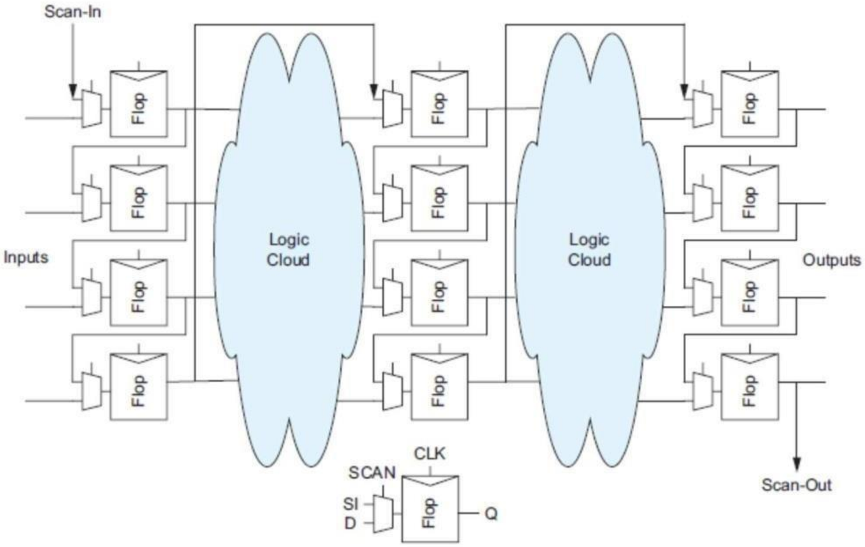
- Partitioninglargesequentialcircuits
- Addingtestpoints
- Addingmultiplexers
- Providingforeasy
- statereset

A technique classified in this category is the use of the bus in a bus-oriented system for test purposes. Each register has been made loadable from the bus and capable of being driven onto the bus.Here,the internallogicvaluesthatexiston adatabus areenabledonto the bus for testing purposes. Frequently, multiplexers can be used to provide alternative signal paths during testing. In CMOS, transmission gate multiplexers provide low area and delay overhead. Anydesignshouldalwayshaveamethodofresettingtheinternalstateofthechipwithina single cycle or at most a few cycles. Apart from making testing easier, this also makes simulation faster as a few cycles are required to initialize the chip.

In general, ad hoc testing techniques represent a bag of tricks developed over the years by designers to avoid the overhead of a systematic approach to testing, as will be described in the next section. While these general approaches are still quite valid, process densities and chip complexities necessitate a structured approach to testing.

SCANBASEDTECHNIQUE

The scan-design strategy for testing has evolved to provide observability and controllability at each register. In designs with scan, the registers operate in one of two modes. In normal mode, they behave as expected.In scan mode, they are connected to form a giant shift register called a scan chain spanning the whole chip. By applying N clock pulses in scan mode, all N bits of state in the system can be shifted out and new N bits of state can be shifted in. Therefore, scan mode gives easy observability and controllability of every register in the system.



Modern scan is based on the use of scan registers, as shown in Figure. The scan register is a D flip-flop preceded by a multiplexer. When the SCAN signal is deasserted, the register behaves as a conventional register, storing data on the D input. When SCAN is asserted,the data is loaded from the SI pin, which is connected in shift register fashion to the previous register Q output in the scan chain. For thecircuit to load thescan chain, SCAN is asserted and CLKispulsedeighttimestoloadthefirsttworanksof4-bitregisterswithdata.SCANis deasserted and CLK is asserted for one cycle to operate the circuit normally with predefined inputs. SCANisthen reasserted and CLK assertedeighttimesto read the stored dataout.At the same time, the new register contents can be shifted in for the next test. Testing proceeds in this manner ofseriallyclockingthedatathroughthescanregister totheright point inthecircuit, running a single system clock cycle and serially clocking the data out for observation. In this scheme, every input to the combinational block can be controlled and every output can be observed.In addition,running a random pattern of 1s and 0sthrough the scan chain can test the chain itself.

Test generation for this type of test architecture can be highly automated. ATPG techniques can be used for the combinational blocks and, as mentioned, the scan chain is easily tested. The prime disadvantage is the area and delay impact of the extra multiplexer in the scan register. Designers (and managers alike) are in widespread agreement thatthiscost ismore than offset by the savings in debug time and production test cost.

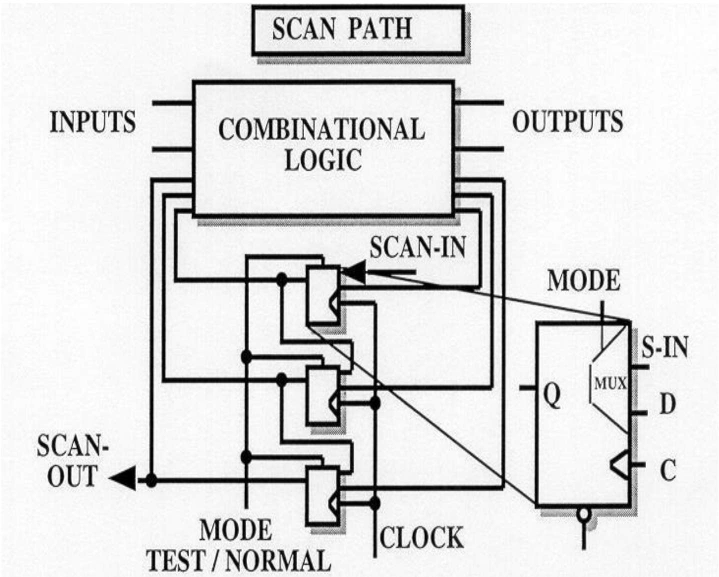
ScanDesignTechniques

Theset of designfor testabilityguidelines presentedaboveis a set of adhoc methods to design random logic in respect with testability requirements. The scan design techniques are a set of structured approaches to design (for testability) the sequential circuits.

The major difficulty in testing sequential circuits is determining the internal state of the circuit. Scandesigntechniques aredirectedat improvingthecontrollabilityandobservabilityof the internal states of a sequential circuit. By this the problem of testing a sequential circuit is reducedtothat oftestinga combinationalcircuit, sincetheinternalstates ofthecircuit areunder control.

8.8.1ScanPath

Thegoal ofthescanpathtechniqueis to reconfigurea sequentialcircuit, for thepurpose of testing, into a combinational circuit. Since a sequential circuit is based on a combinational circuit andsome storage elements, thetechniqueof scanpathconsists inconnectingtogether all thestorage elements to forma long serial shift register. Thus theinternal stateof the circuit can be observed and controlled by shifting (scanning) out the contents of the storage elements. The shift register is then called a scan path.



Thestorage elements can either beD, J-K, or R-S types of flip-flops, but simple latches cannot beused inscanpath. However, thestructureof storageelements is slightly different than classical ones. Generallytheselectionof theinput sourceis achieved usinga multiplexer onthe data input controlled by an external mode signal. This multiplexer is integrated into the D-flip- flop, in our case; the D-flip-flop is then called MD-flip-flop (multiplexed-flip-flop).

The sequential circuit containing a scan path has two modes of operation : a normal mode and a test mode which configure the storage elements in the scan path.

In the normal mode, the storage elements are connected to the combinational circuit, in theloops of the global sequential circuit, which is considered then as a finitestate machine.

In the testmode, the loops are broken and the storage elements are connected togetheras aserial shift register(scan path),receiving thesame clock signal.The inputof thescan path is called scan-in and the output scan-out. Several scan paths can be implemented in one same complex circuit if it is necessary, though having several scan-in inputs and scan-out outputs.

A large sequential circuit can be partitioned into sub-circuits, containing combinational sub-circuits, associated with one scan path each. Efficiency of the test pattern generation for a combinational sub-circuit is greatly improved by partitioning, since its depth is reduced.

Before applying test patterns, the shift register itself has to be verified by shifting in all ones i.e. 111...11, or zeros i.e. 000...00, and comparing.

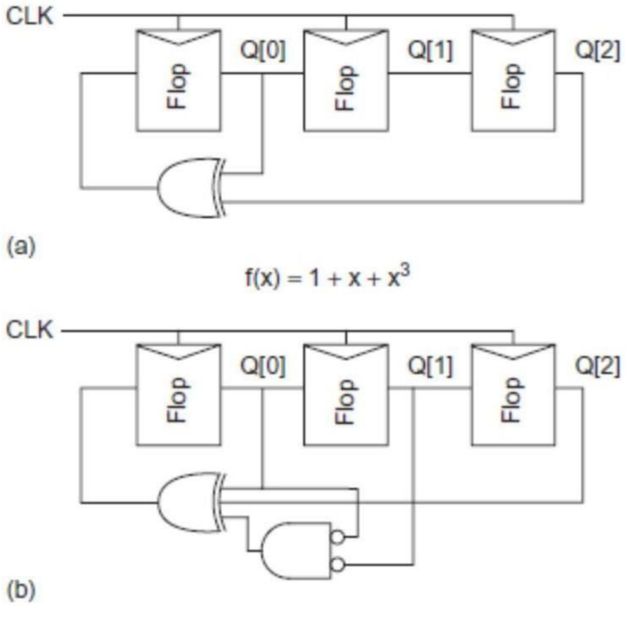
Themethod oftestingacircuitwiththescanpathis asfollows:

1. Settestmodesignal,flip-flopsacceptdatafrominputscan-in

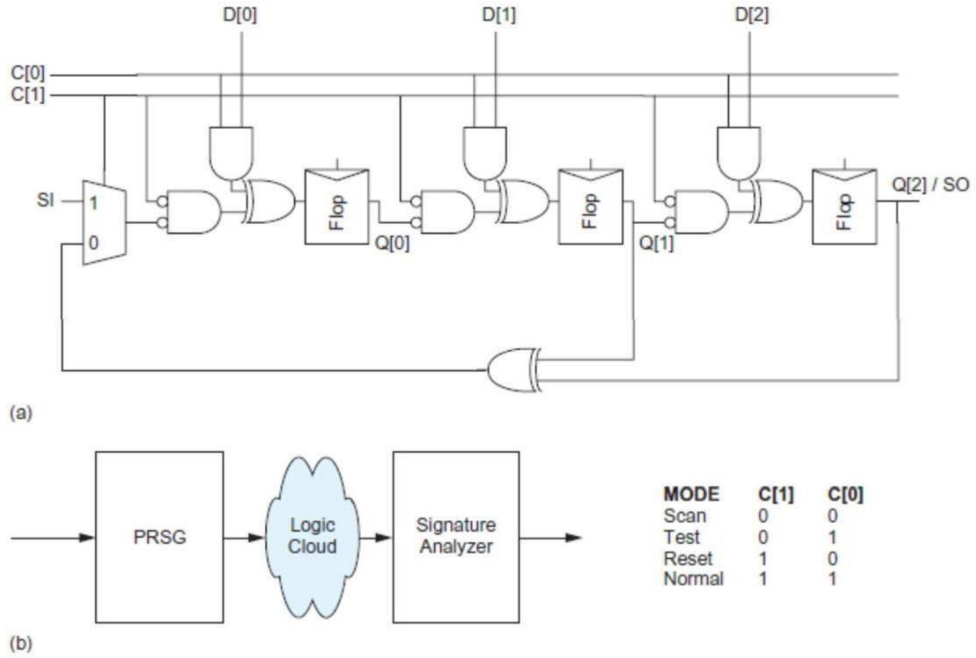
2. Verifythescanpathbyshiftinginandouttestdata
3. Settheshift registertoaninitialstate
4. Applyatestpatterntotheprimaryinputs ofthecircuit
5. Setnormalmode, thecircuitsettlesandcanmonitortheprimaryoutputsofthecircuit
6. Activatethecircuit clockforonecycle
7. Returntotestmode
8. Scanoutthecontents oftheregisters,simultaneouslyscaninthenextpattern

SELF-TESTAPPROACHES:BUILT-INSELF-TEST(BIST)

Self-test andbuilt-intest techniques, astheir names suggest, relyonaugmentingcircuits to allow them to perform operations upon themselves that prove correct operation. These techniques add area to the chip for the test logic, but reduce the test time required and thus can lower the overall system cost. [Stroud02] offers extensive coverage of the subject from the implementer's perspective.

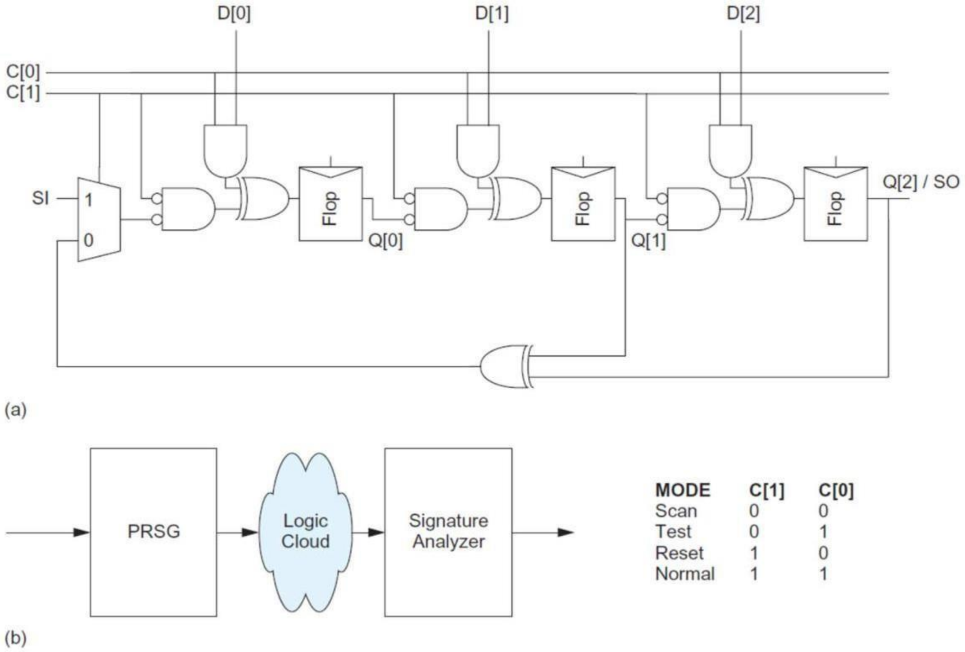


One method of testing a module is to use signature analysis or cyclic redundancy checking. This involves using a pseudo-random sequence generator (PRSG) toproduce the input signals for a section of combinational circuitry and a signature analyzer to observe the output signals. A PRSG of length n is constructed from a linear feedback shift register (LFSR), which in turn is made of n flip-flops connected in a serial fashion, as shown in Figure (a). The XOR of particular outputs are fed back to the input of the LFSR. An n-bit LFSR will cycle through 2n–1 states before repeating the sequence.LFSRs are discussedfurther in SectionThey are described by a characteristic polynomial indicating which bits are fed back.A complete feedback shift register(CFSR), shown in Figure (b), includes the zero state that may be required in some testsituations.Ann-bitLFSRisconverted to ann-bitCFSR by adding an n – 1 input NOR gate connected to all but the last bit. When in state 0…01, the next state is 0…00. When in state 0…00, the next state is 10…0. Otherwise, the sequence is the same. Alternatively, the bottom n bits of an n + 1-bit LFSR can be used to cycle through the all zeros state without the delay of the NOR gate.



A signature analyzer receives successive outputs of a combinational logic block and produces a syndrome that is a function of these outputs. The syndrome is reset to 0, and then XORed with the output on each cycle. The syndrome is swizzled each cycle so that afault in one bit is unlikely to cancel itself out. At the end of a test sequence, the LFSR contains the syndromethatisafunctionofallpreviousoutputs.Thiscanbecomparedwiththecorrect syndrome (derived by running a test program on the good logic) to determine whetherthe circuit is good or bad. If the syndrome contains enough bits, it is improbable that a defective circuit will produce the correct syndrome.

The combination of signature analysis and the scan technique creates a structure known as BIST—for Built-In Self-Test or BILBO—for Built-In Logic Block Observation The 3-bit BIST register shown in Figure is a scannable, resettable register that also can serve as a pattern generator and signature analyzer. C[1:0] specifies the mode of operation.In the resetmode(10), all the flip-flops are synchronously initialized to 0. In normal mode (11), the flip-flops behave normally with their D input and Q output. In scan mode (00), the flip-flops are configured as a 3-bit shift register between SI and SO. Note that there is an inversion between each stage. In test mode (01), the register behaves as a pseudo- random sequence generator or signatureanalyzer. IfalltheD inputs areheld low, theQ outputs loop througha pseudo-random bit sequence, which can serve as the input to the combinational logic. If the D inputs are taken from the combinational logic output, they are swizzled with the existing state to produce the syndrome. Insummary, BIST is performedbyfirst resettingthesyndrome inthe output register. Then both registers are placed in the test mode to produce the pseudo-random inputs and calculate the syndrome. Finally, the syndrome is shifted out through the scan chain.



MODE	C[1]	C[0]
Scan	0	0
Test	0	1
Reset	1	0
Normal	1	1

Various companies have commercial design aid packages that automatically replace ordinary registers with scannable BIST registers, check the fault coverage, and generate scripts for production testing. As an example, on a WLAN modem chip comprising roughly 1 million gates, a full at-speed test takes under a second with BIST. This comes with roughly a 7.3% overhead in the core area (but actually zero because the design was pad limited) and a 99.7% fault coverage level. The WLANmodem parts designed in this way were fully tested in less than ten minutes on receipt of first silicon. This kind of test method is incredibly valuable for productivity in manufacturing test generation.

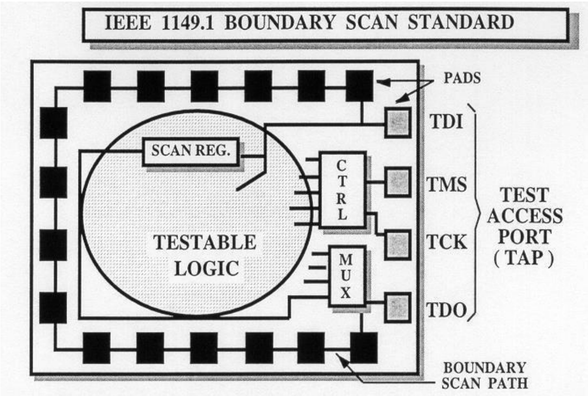
MemoryBIST

On many chips, memories account for the majority of the transistors. A robust testing methodology must be applied to provide reliable parts. In a typical MBISTscheme,multiplexers are placed on the address, data, and control inputs for the memory to allow direct access during test. During testing, a state machine uses these multiplexers to directly write a checkerboard pattern of alternating 1s and 0s. The data is read back, checked, then the inverse patternisalso applied and checked.ROMtestingiseven simpler:The contentsare read outto a signature analyzer to produce a syndrome.

BoundaryScanTest(BST)

Boundary Scan Test (BST) is a technique involving scan path andself-testing techniques to resolve the problem of testing boards carrying VLSI integrated circuits and/or surface mounted devices (SMD).

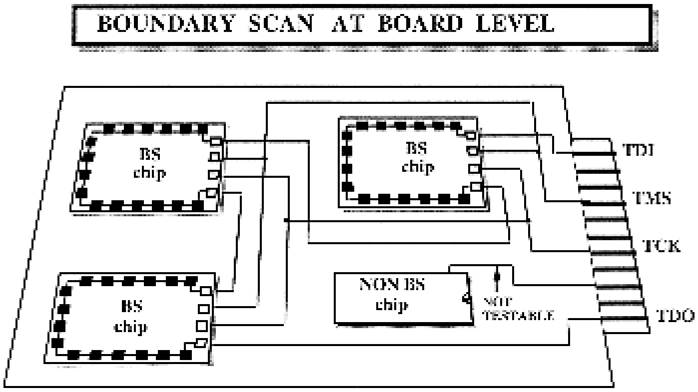
Printed circuit boards (PCB) are becoming very dense and complex, especially with SMD circuits, that most test equipment cannot guaranteea good fault coverage.



BST consists in placing a scan path (shift register) adjacent to each component pin and to interconnect the cells in order to form a chain around the border of the circuit. The BST circuits contained on one board are then connected together to form a single path through the board.

Theboundaryscanpathisprovidedwithserialinputandoutputpadsandappropriate clockpads whichmakeit possibleto:

- Testtheinterconnectionsbetweenthevariouschip
- Delivertestdatatothechipsonboardforself-testing
- Testthechipsthemselveswithinternalself-test



TheadvantagesofBoundaryscantechniquesareasfollows:

- NoneedforcomplextestersinPCBtesting
- Testengineersworkissimplifiedandmoreefficient
- Timetospendontestpatterngenerationandapplicationisreduced
- Faultcoverageisgreatlyincreased.

BS Techniques are grouped by the IEEE Standard Organization in a"standard test access portandboundaryscanarchitecture", namelyIEEEP1149.1-1990.TheJoint Test Action Group (JTAG), formed basically in 1986 at Philips, is an international committee composed of IC manufacturers who have set the technical development of the IEEE P1149 standard and promoted its use by all sectors of electronics industry.

TheIEEE 1149 is a family of overall testability bus standards, defined by the Joint Test Action Group (JTAG), formed basically in 1986 at Philips. JTAG is an international committee composed of European and American IC manufacturers. The "standard Test Access Port and Boundary Scan architecture", namely IEEE P1149.1 accepted by the IEEE standard committee in February1990, is the first one of this family. Several other ongoing standards are developed and suggested as drafts to the technical committee of the IEEE 1149 standard in order to promote their use by all sectors of electronics industry.

****