VLSI CIRCUITS AND DESIGN

Subject Code	:	10EE764	IA Marks	:	25
No. of Lecture Hrs./ Week	:	04	Exam Hours	:	03
Total No. of Lecture Hrs.	:	52	Exam Marks	:	100

PART - A

UNIT - 1

A REVIEW OF MICROELECTRONIC 3 AND AN INTRODUCTION TO MOS TECHNOLOGY: Introduction to integrated circuit technology, Production of E-beam masks. Introduction, VLSI technologies, MOS transistors, fabrication, thermal aspects, production of E-beam masks.

6 Hours

UNIT - 2

BASIC ELECTRICAL PROPERTIES OF MOS AN BICMOS CIRCUIT: Rain to source current Ids versus

Vds relationships-BICMOS latch up susceptibility. MOS transistor characteristics, figure of merit,

pass transistor NMOS and COMS inverters, circuit model, latch up.

8 Hours

UNIT - 3

MOS AND BICMOS CIRCUIT DESIGN PROCESSES: Mass layers, strick diagrams, design, symbolic diagrams

8 Hours

UNIT - 4

BASIC CIRCUIT CONCEPTS: Sheet resistance, capacitance layer inverter delays, wiring capacitance, choice of layers.

6 Hours

PART - B

UNIT - 5

SCALING OF MOS CIRCUITS: Scaling model and scaling factors- Limit due to current density.

8 Hours

UNIT - 6

SUBSYSTEM DESIGN AND LAYOUT: Some architecture issues- other systems considerations. Examples of structural design, clocked sequential circuits

8 Hours

UNIT - 7

SUBSYSTEM DESIGN PROCESSES: Some general considerations, an Illustration of design process, Observations. **4 Hours**

UNIT - 8

ILLUSTRATION OF THE DESIGN PROCESS: Observation on the design process, Regularity Design of an ALU subsystem. Design of 4-bit adder, implementing ALU functions. **4 Hours**

TEXT BOOKS:

1. **"Basic VLSI Design" -3rd** Edition, PHI
2. **"Fundamentals of Modern VLSI Devices"**-Yuan Taun Tak H Ning Cambridge Press, South Asia Edition 2003,
3. **"ModernVLSI Design Wayne wolf", Pearson Education Inc. 3rd edition"**-Wayne wolf 2003.

	Some general considerations.	
	an Illustration of design process.	
	Observations.	
8	**UNIT – 8 Illustration of the Design Process**	**196-210**
	Observation on the design process.	
	Regularity Design of an ALU subsystem.	
	Design of 4-bit adder.	
	Implementing ALU functions.	

Unit 1
Basic MOS Technology

Transistor was first invented by William.B.Shockley, Walter Brattain and John Bardeen of Bell Labratories. In 1961, first IC was introduced.

Levels of Integration:-

i) SSI:- (10-100) transistors => Example: Logic gates ii) MSI:- (100-1000) => Example: counters
iii) LSI:- (1000-20000) => Example:8-bit chip
iv) VLSI:- (20000-1000000) => Example:16 & 32 bit up
v) ULSI:- (1000000-10000000) => Example: Special processors, virtual reality machines, smart sensors

Moore ’s Law:-

“The number of transistors embedded on the chip doubles after every one and a half years.” The number of transistors is taken on the y-axis and the years in taken on the x- axis. The diagram also shows the speed in MHz. the graph given in figure also shows the variation of speed of the chip in MHz.

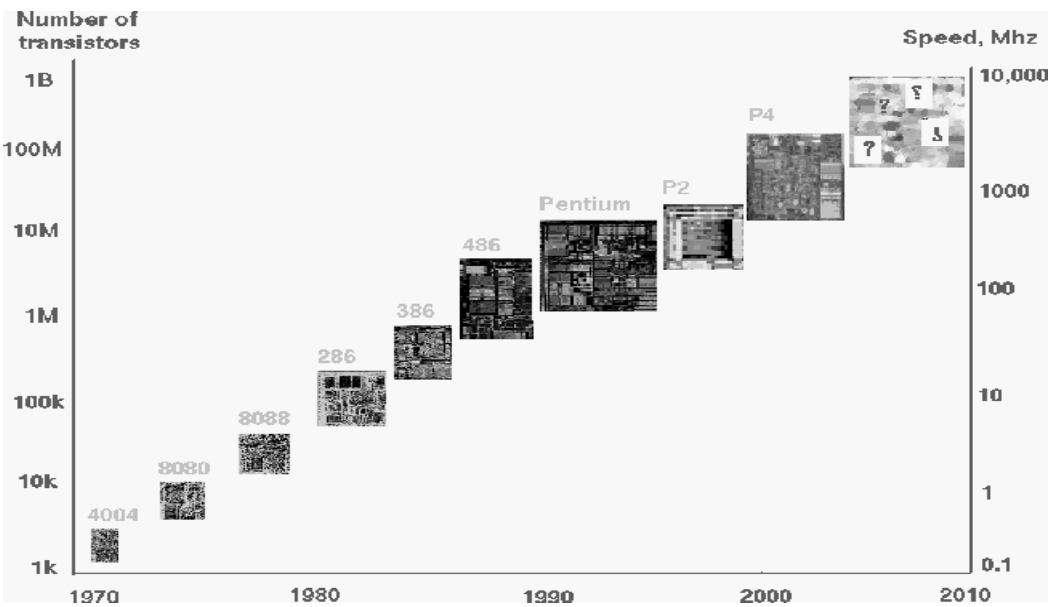


Figure 1. Moore’s law

The graph in figure2 compares the various technologies available in ICs.

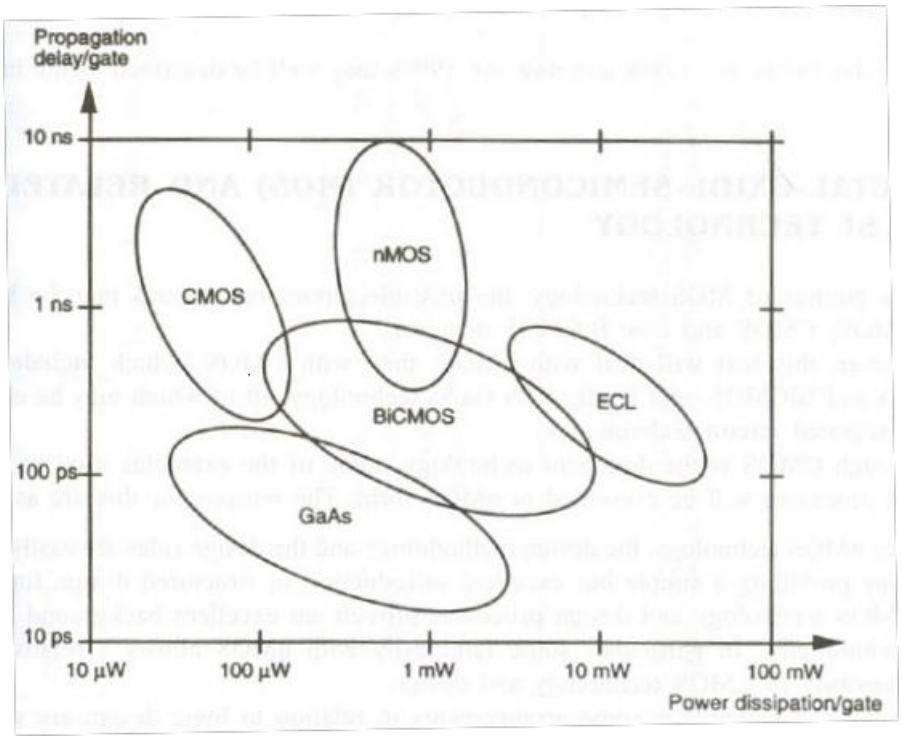


Figure 2.Comparison of available technologies

From the graph we can conclude that GaAs technology is better but still it is not used because of growing difficulties of GaAs crystal. CMOS looks to be a better option compared to nMOS since it consumes a lesser power. BiCMOS technology is also used in places where high driving capability is required and from the graph it confirms that, BiCMOS consumes more power compared to CMOS.

Levels of Integration:-

i) Small Scale Integration:- (10-100) transistors => Example: Logic gates ii) Medium Scale Integration:- (100-1000) => Example: counters

iii) Large Scale Integration:- (1000-20000) => Example:8-bit chip

iv) Very Large Scale Integration:- (20000-1000000) => Example:16 & 32 bit up

v) Ultra Large Scale Integration:- (1000000-10000000) => Example: Special processors, virtual reality machines, smart sensors

Basic MOS Transistors:

Why the name MOS?

We should first understand the fact that why the name Metal Oxide Semiconductor transistor, because the structure consists of a layer of Metal (gate), a layer of oxide (Sio2) and a layer of semiconductor. Figure 3 below clearly tell why the name MOS.

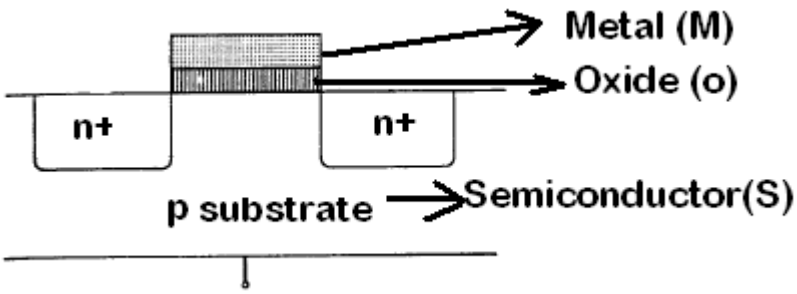


Figure 3.cross section of a MOS structure

We have two types of FETs. They are Enhancement mode and depletion mode transistor. Also we have PMOS and NMOS transistors.

In **Enhancement mode transistor** channel is going to form after giving a proper positive gate voltage. We have NMOS and PMOS enhancement transistors.

In **Depletion mode transistor** channel will be present by the implant. It can be removed by giving a proper negative gate voltage. We have NMOS and PMOS depletion mode transistors.

N-MOS enhancement mode transistor:-

This transistor is normally off. This can be made ON by giving a positive gate voltage. By giving a +ve gate voltage a channel of electrons is formed between source drain.

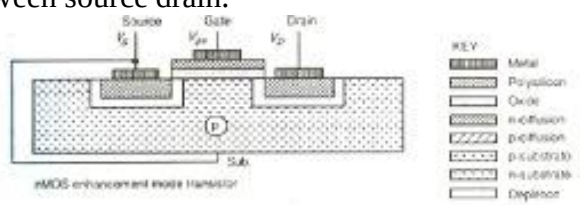


Fig 5. Nmos Enhancement transistor

P-Mos enhancement mode transistors:-

This is normally on. A Channel of Holes can be performed by giving a –ve gate voltage. In P-Mos current is carried by holes and in N-Mos its by electrons. Since the mobility is of holes less than that of electrons P-Mos is slower.

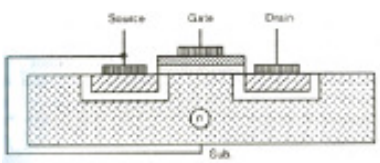


Fig 6. Pmos Enhancement transistor

N-MOS depletion mode transistor:-

This transistor is normally ON, even with Vgs=0. The channel will be implanted while fabricating, hence it is normally ON. To cause the channel to cease to exist, a –ve voltage must be applied between gate and source.

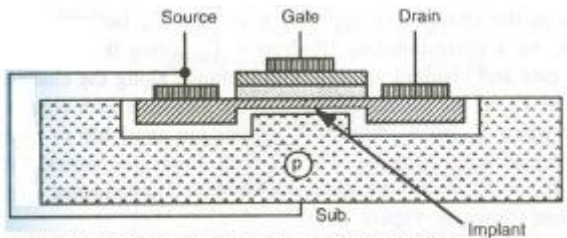
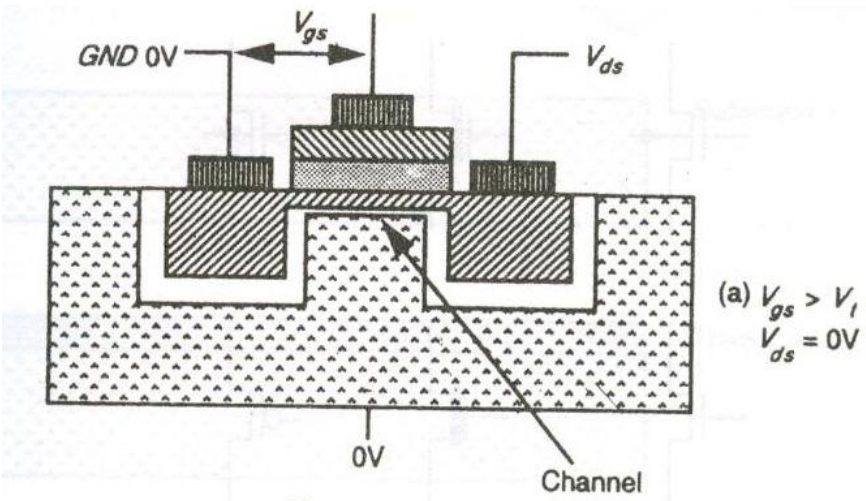
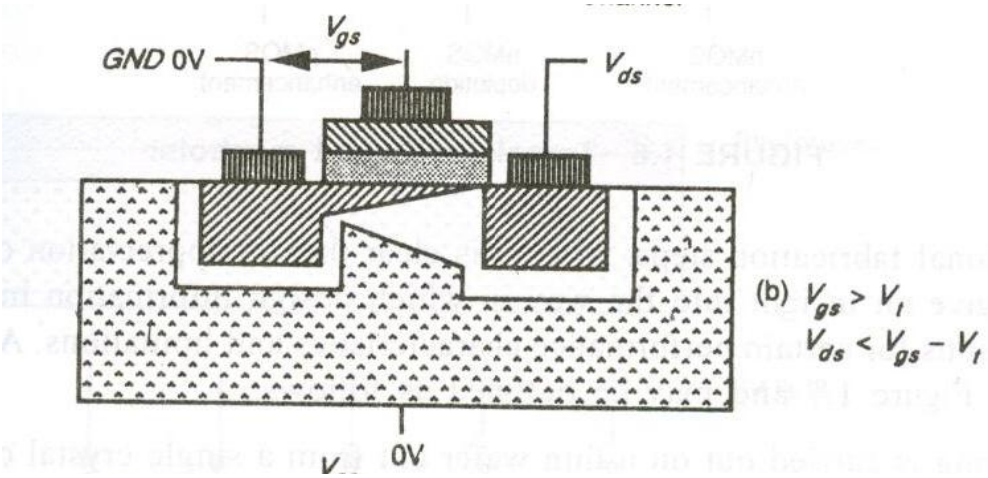


Fig. 7 Nmos Depletion mode transistor

NOTE: Mobility of electrons is 2.5 to 3 times faster than holes. Hence P-MOS devices will have more resistance compared to NMOS.

Enhancement mode Transistor action:-





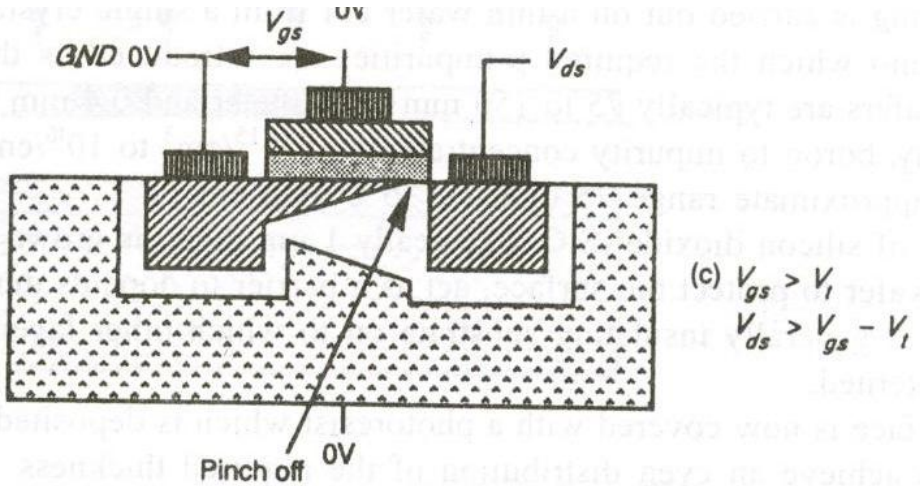


Figure 8(a)(b)(c) Enhancement mode transistor with different Vds values

To establish the channel between the source and the drain a minimum voltage (V_t) must be applied between gate and source. This minimum voltage is called as "Threshold Voltage". The complete working of enhancement mode transistor can be explained with the help of diagram a, b and c.

a) $V_{gs} > V_t$
$V_{ds} = 0$

Since $V_{gs} > V_t$ and $V_{ds} = 0$ the channel is formed but no current flows between drain and source.

b) Vgs > Vt
Vds < Vgs - Vt

This region is called the non-saturation Region or linear region where the drain current increases linearly with Vds. When Vds is increased the drain side becomes more reverse biased(hence more depletion region towards the drain end) and the channel starts to pinch. This is called as the pinch off point.

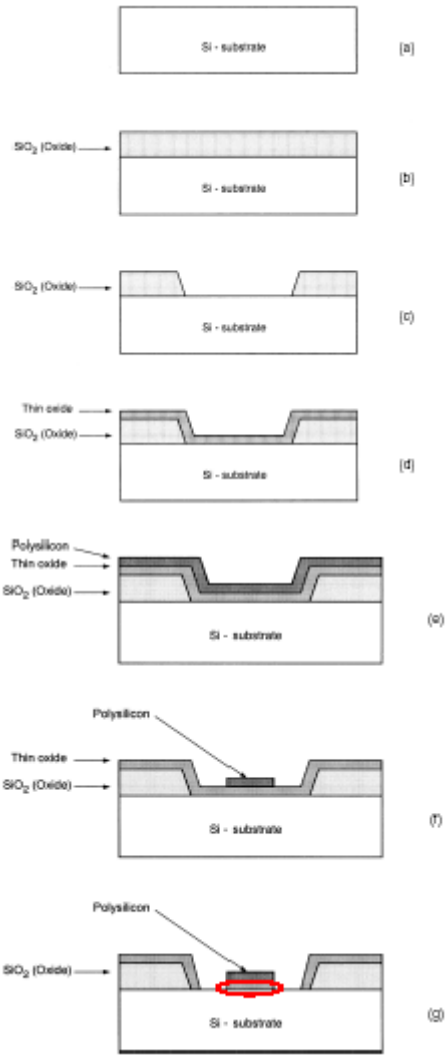
c) Vgs > Vt
Vds > Vgs - Vt

This region is called Saturation Region where the drain current remains almost constant. As the drain voltage is increased further beyond (Vgs-Vt) the pinch off point starts to move from the drain end to the source end. Even if the Vds is increased more and more, the increased voltage gets dropped in the depletion region leading to a constant current.
The typical threshold voltage for an enhancement mode transistor is given by Vt = 0.2 * Vdd.

Depletion mode Transistor action:-

We can explain the working of depletion mode transistor in the same manner, as that of the enhancement mode transistor only difference is, channel is established due to the implant even when Vgs = 0 and the channel can be cut off by applying a –ve voltage between the gate and source. Threshold voltage of depletion mode transistor is around

NMOS
Fabrication:



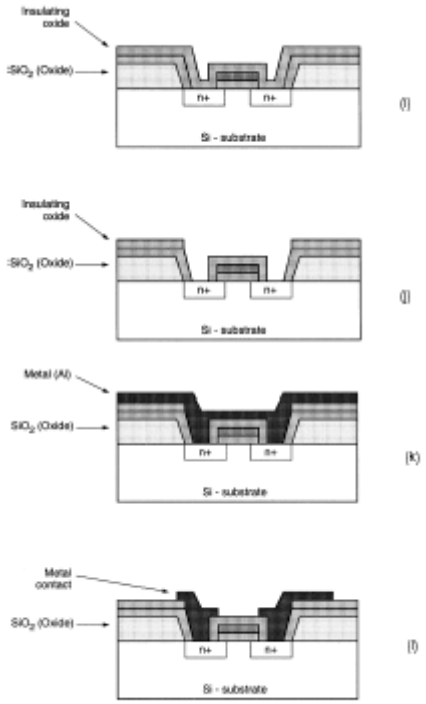


Figure 9 NMOS Fabrication process steps

The process starts with the oxidation of the silicon substrate (Fig. 9(a)), in which a relatively thick silicon dioxide layer, also called field oxide, is created on the surface (Fig. 9(b)). Then, the field oxide is selectively etched to expose the silicon surface on which the MOS transistor will be created (Fig. 9(c)). Following this step, the surface is covered with a thin, high-quality oxide layer, which will eventually form the gate oxide of the MOS transistor (Fig. 9(d)). On top of the thin oxide, a layer of polysilicon (polycrystalline silicon) is deposited (Fig. 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 resistivity of polysilicon can be reduced, however, by doping it with impurity atoms.

After deposition, the polysilicon layer is patterned and etched to form the interconnects and the MOS transistor gates (Fig. 9(f)). The thin gate oxide not covered by polysilicon is also etched away, which exposes the bare silicon surface on which the source and drain junctions are to be formed (Fig. 9(g)). The entire silicon surface is then doped with a high concentration of impurities, either through diffusion or ion implantation (in this case with donor atoms to produce n-type doping). Figure 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 impurity doping also penetrates the polysilicon on the surface, reducing its resistivity. Note that the polysilicon gate, which is patterned before doping actually defines the precise location of the channel region and, hence, the location of the source and the drain regions. Since this procedure allows very precise positioning of the two regions relative to the gate, it is also called the self-aligned

process. Once the source and drain regions are completed, the entire surface is again covered with an insulating layer of silicon dioxide (Fig. 9 (i)). The insulating oxide layer is then patterned in order to provide contact windows for the drain and source junctions (Fig. 9 (j)). The surface is covered with evaporated aluminum which will form the interconnects (Fig. 9 (k)). Finally, the metal layer is patterned and etched, completing the interconnection of the MOS transistors on the surface (Fig. 9 (l)). Usually, a second (and third) layer of metallic interconnect can also be added on top of this structure by creating another insulating oxide layer, cutting contact (via) holes, depositing, and patterning the metal.

CMOS fabrication: When we need to fabricate both nMOS and pMOS transistors on the same substrate we need to follow different processes. The three different processes are,P-well process ,N-well process and Twin tub process.

P-WELL PROCESS:

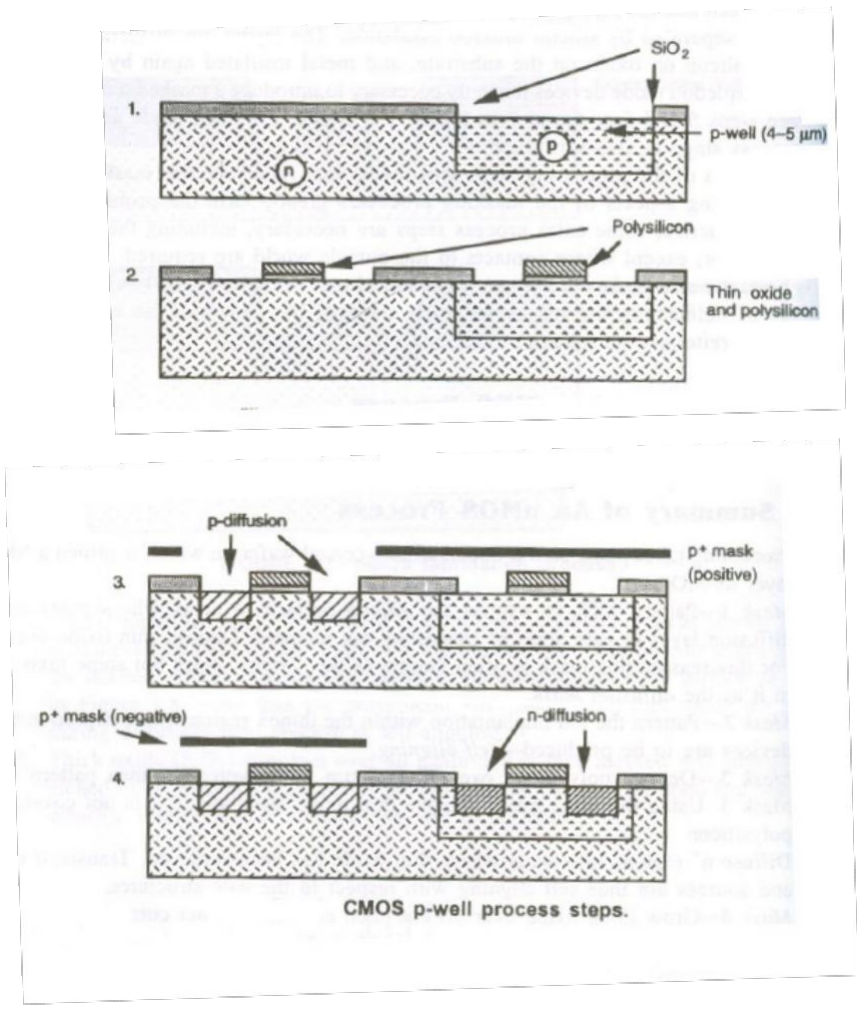


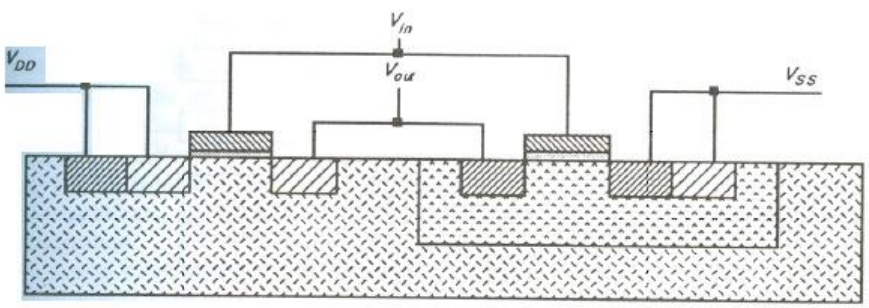
Figure 10 CMOS Fabrication (P-WELL) process steps

The p-well process starts with a n type substrate. The n type substrate can be used

to implement the pMOS transistor, but to implement the nMOS transistor we need to provide a p-well, hence we have provided he place for both n and pMOS transistor on the same n-type substrate.

Mask sequence.
Mask 1:
Mask 1 defines the areas in which the deep p-well diffusion takes place.
Mask 2:
It defines the thin oxide region (where the thick oxide is to be removed or stripped and thin oxide grown)
Mask 3:
It's used to pattern the polysilicon layer which is deposited after thin oxide.
Mask 4:
A p+ mask (anded with mask 2) to define areas where p-diffusion is to take place.
Mask 5:
We are using the –ve form of mask 4 (p+ mask) It defines where n-diffusion is to take place.
Mask 6:
Contact cuts are defined using this mask.
Mask 7:
The metal layer pattern is defined by this mask.
Mask 8:
An overall passivation (overglass) is now applied and it also defines openings for accessing pads.

The cross section below shows the CMOS pwell inverter.
.



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

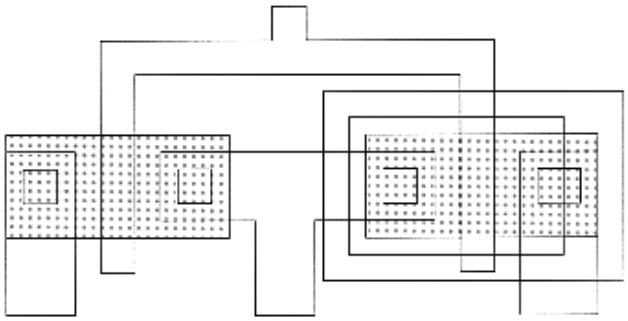
Figure 11 CMOS inverter (P-WELL)

N-WELL PROCESS:

In the following figures, some of the important process steps involved in the fabrication of a CMOS inverter will be shown by a top view of the lithographic masks and a cross-sectional view of the relevant areas.

The n-well CMOS process starts with a moderately doped (with impurity concentration typically less than 10^{15} cm-3) p-type silicon substrate. Then, an initial

oxide layer is grown on the entire surface. The first lithographic mask defines the n-well region. Donor atoms, usually phosphorus, are implanted through this window in the oxide. Once the n-well is created, the active areas of the nMOS and pMOS transistors can be defined. Figures 12.1 through 12.6 illustrate the significant milestones that occur during the fabrication process of a CMOS inverter.



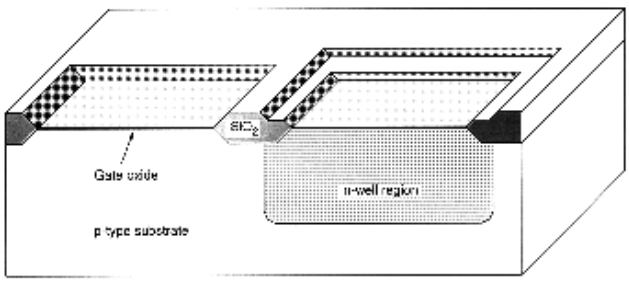


Figure-12.1: Following the creation of the n-well region, a thick field oxide is grown in the areas surrounding the transistor active regions, and a thin gate oxide is grown on top of the active regions. The thickness and the quality of the gate oxide are two of the most critical fabrication parameters, since they strongly affect the operational characteristics of the MOS transistor, as well as its long-term reliability.

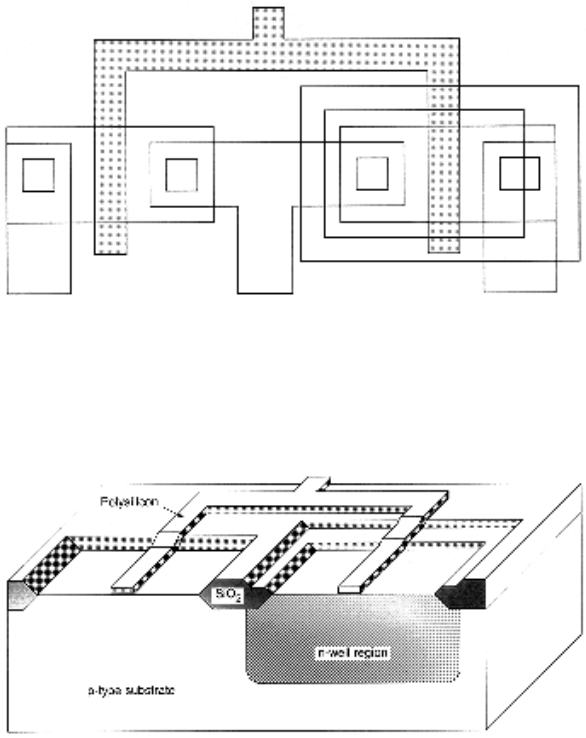


Figure-12.2: The polysilicon layer is deposited using chemical vapor deposition (CVD) and patterned by dry (plasma) etching. The created polysilicon lines will function as the gate electrodes of the nMOS and the pMOS transistors and their interconnects. Also, the polysilicon gates act as self-aligned masks for the source and drain implantations that follow this step.

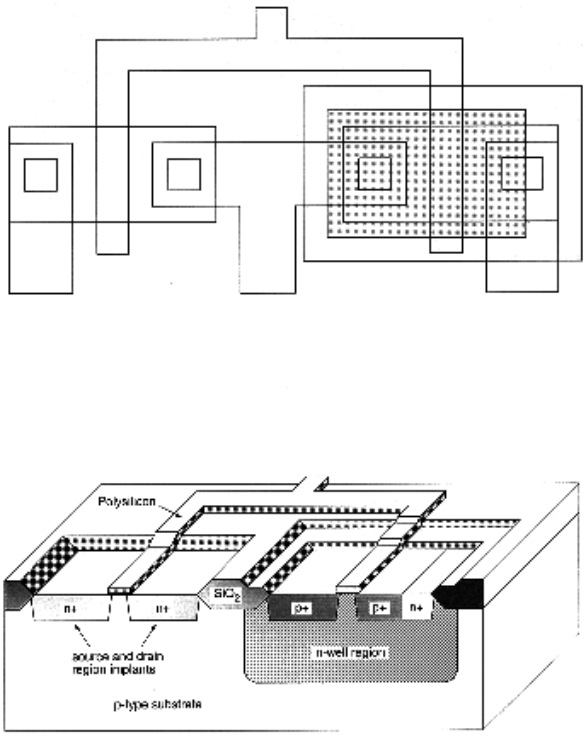


Figure-12.3: Using a set of two masks, the n+ and p+ regions are implanted into the substrate and into the n- well, respectively. Also, the ohmic contacts to the substrate and to the n-well are implanted in this process step.

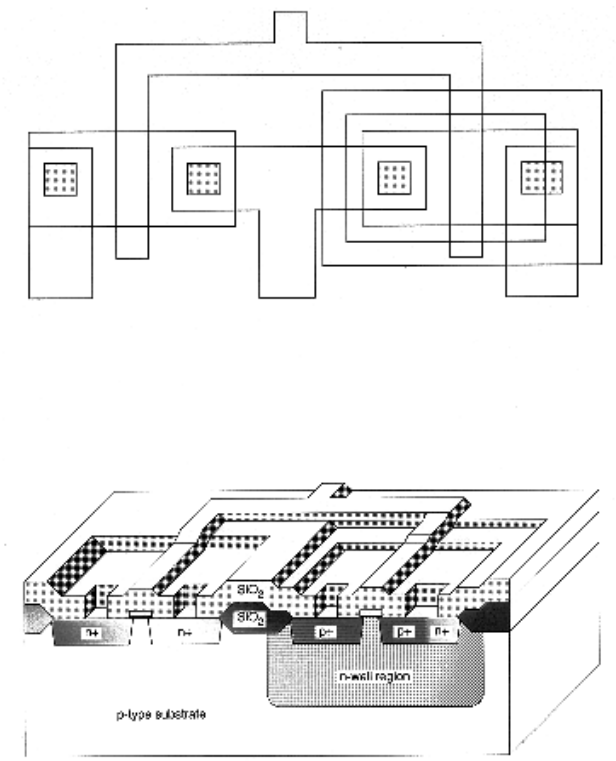
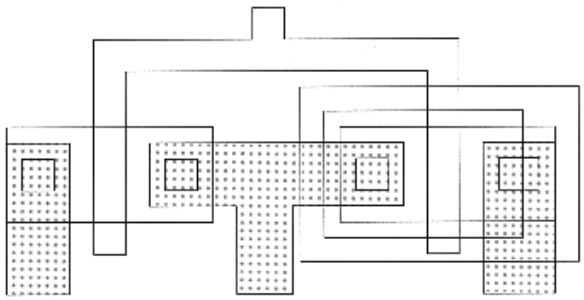


Figure-12.4: An insulating silicon dioxide layer is deposited over the entire wafer using CVD. Then, the contacts are defined and etched away to expose the silicon or polysilicon contact windows. These contact windows are necessary to complete the circuit interconnections using the metal layer, which is patterned in the next step.



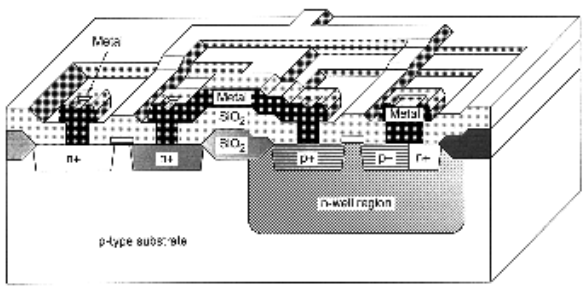


Figure-12.5: Metal (aluminum) is deposited over the entire chip surface using metal evaporation, and the metal lines are patterned through etching. Since the wafer surface is non-planar, the quality and the integrity of the metal lines created in this step are very critical and are ultimately essential for circuit reliability.

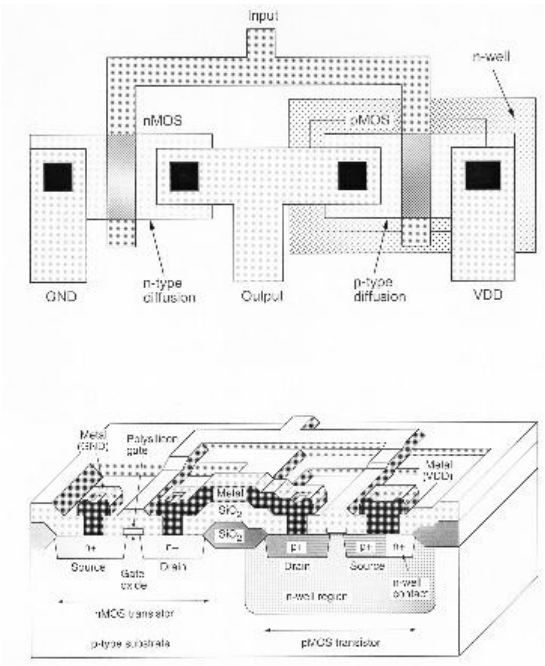


Figure-12.6: The composite layout and the resulting cross-sectional view of the chip, showing one nMOS and one pMOS transistor (built-in n-well), the polysilicon and metal interconnections. The final step is to deposit the passivation layer (for protection) over the chip, except for wire-bonding pad areas.

Twin-tub process:

Here we will be using both p-well and n-well approach. The starting point is a n-type material and then we create both n-well and p-well region. To create the both well we first go for the epitaxial process and then we will create both wells on the same substrate.

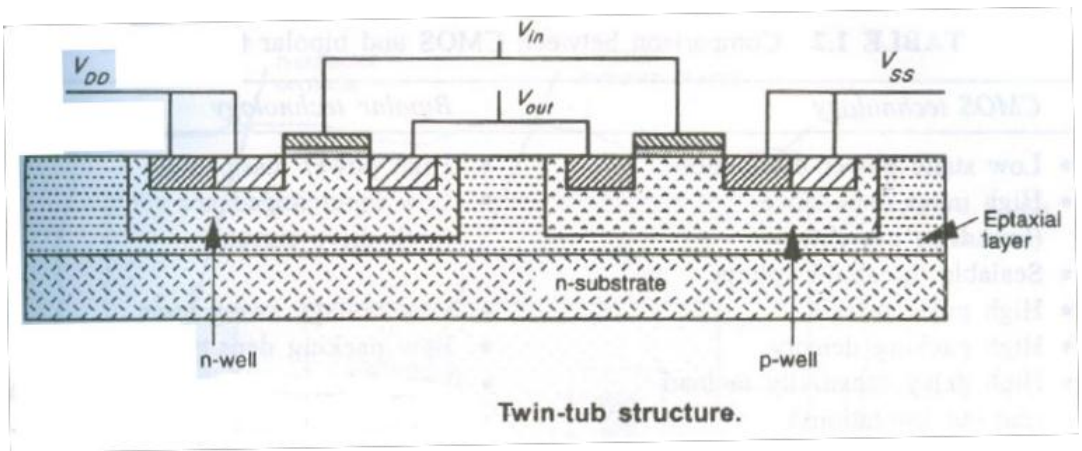


Figure 13 CMOS twin-tub inverter

NOTE: Twin tub process is one of the solutions for latch-up problem.

Bi-CMOS technology: - (Bipolar CMOS)

The driving capability of MOS transistors is less because of limited current sourcing and sinking capabilities of the transistors. To drive large capacitive loads we can think of Bi-Cmos technology.

This technology combines Bipolar and CMOS transistors in a single integrated circuit, by retaining benefits of bipolar and CMOS, BiCMOS is able to achieve VLSI circuits with speed-power-density performance previously unattainable with either technology individually.

Characteristics of CMOS Technology

- Lower static power dissipation
- Higher noise margins
- Higher packing density – lower manufacturing cost per device
- High yield with large integrated complex functions
- High input impedance (low drive current)
- Scaleable threshold voltage
- High delay sensitivity to load (fan-out limitations)
- Low output drive current (issue when driving large capacitive loads)
- Low transconductance, where transconductance, gm α Vin
- Bi-directional capability (drain & source are interchangeable)
- A near ideal switching device

Characteristics of Bipolar Technology

- Higher switching speed
- Higher current drive per unit area, higher gain
- Generally better noise performance and better high frequency characteristics
- Better analogue capability
- Improved I/O speed (particularly significant with the growing importance of package limitations in high speed systems).
- high power dissipation
- lower input impedance (high drive current)
- low voltage swing logic
- low packing density
- low delay sensitivity to load
- high gm (gm α Vin)
- high unity gain band width (ft) at low currents
- essentially unidirectional

From the two previous paragraphs we can get a comparison between bipolar and CMOS technology.

The diagram given below shows the cross section of the BiCMOS process which uses an npn transistor.

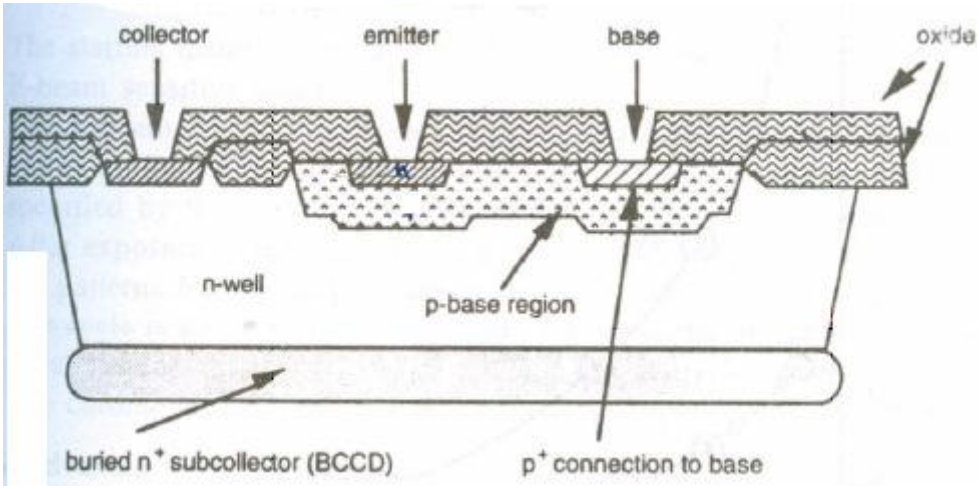


Figure 14 Cross section of BiCMOS process

The figure below shows the layout view of the BiCMOS process.

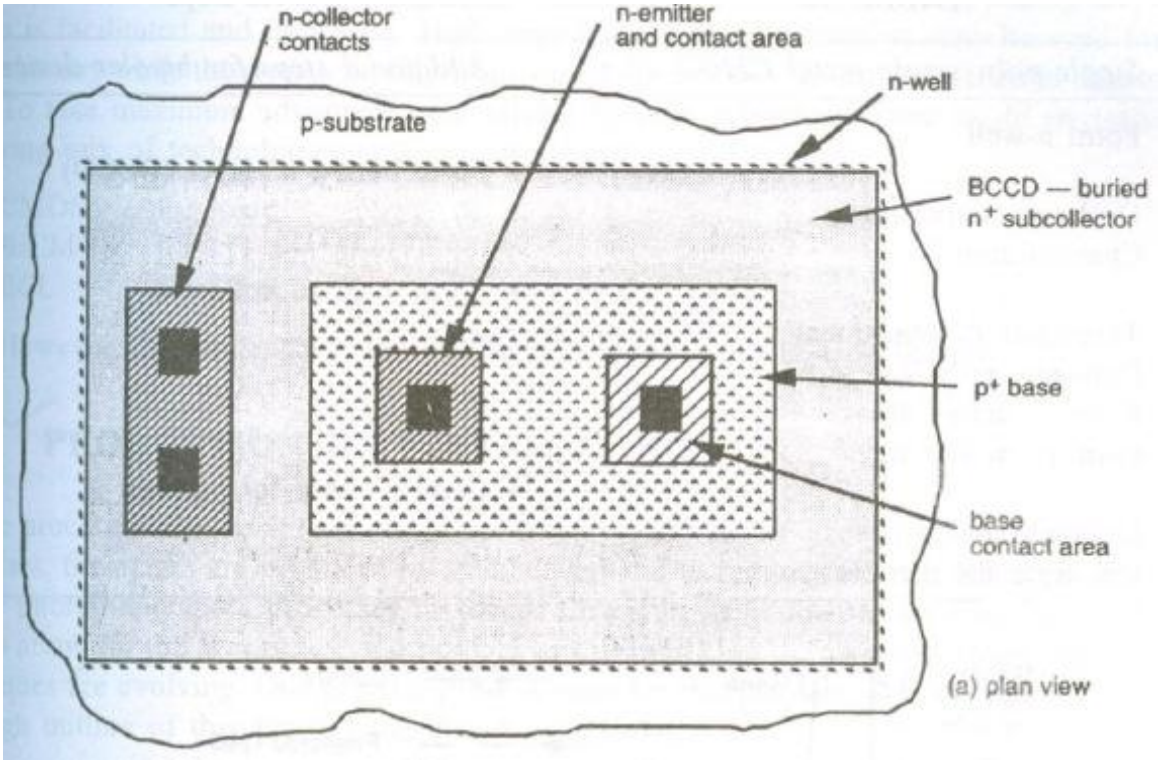


Fig.15. Layout view of BiCMOS process

The graph below shows the relative cost vs. gate delay.

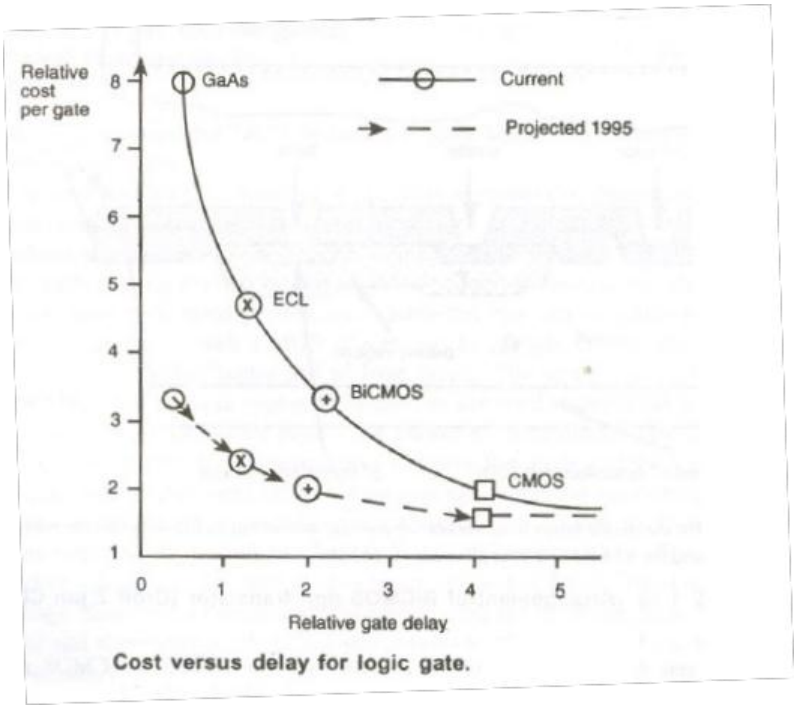


Fig.16. cost versus delay graph

Production of e-beam masks:

In this topic we will understand how we are preparing the masks using e-beam technology. The following are the steps in production of e-beam masks.

- Starting materials is chromium coated glass plates which are coated with e-beam sensitive resist.
- E-beam machine is loaded with the mask description data.
- Plates are loaded into e-beam machine, where they are exposed with the patterns specified by mask description data.
- After exposure to e-beam, plates are introduced into developer to bring out patterns.
- The cycle is followed by a bake cycle which removes resist residue.
- The chrome is then etched and plate is stripped of the remaining e-beam resist.

We use two types of scanning, Raster scanning and vector scanning to map the pattern on to the mask.

In raster type, e-beam scans all possible locations and a bit map is used to turn the e-beam on and off, depending on whether the particular location being scanned is to be exposed or not.

Advantages e-beam masks:

Tighter layer to layer registration;
Small feature sizes.

UNIT - 2

BASIC ELECTRICAL PROPERTIES OF MOS AN BICMOS CIRCUIT: Rain to source current Ids versus Vds relationships-BICMOS latch up susceptibility. MOS transistor characteristics, figure of merit, pass transistor NMOS and COMS inverters, circuit model, latch up.

Introduction

The present chapter first develops the fundamental physical characteristics of the MOS transistor, in which the electrical currents and voltages are the most important quantities. The link between physical design and logic networks can be established. Figure 2.1 depicts various symbols used for the MOS transistors. The symbol shown in Figure 2.1(a) is used to indicate only switch logic, while that in Figure 2.1(b) shows the substrate connection.

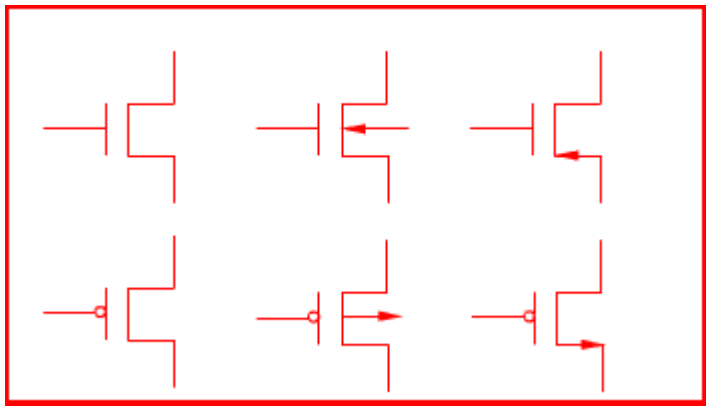


Figure 2.1 Various symbols for MOS transistors

This chapter first discusses about the basic electrical and physical properties of the Metal Oxide Semiconductor (MOS) transistors. The structure and operation of the nMOS and pMOS transistors are addressed, following which the concepts of threshold voltage and body effect are explained. The current-voltage equation of a MOS device for different regions of operation is next established.

It is based on considering the effects of external bias conditions on charge distribution in MOS system and on conductance of free carriers on one hand, and the fact that the current flow depends only on the majority carrier flow between the two device terminals. Various second-order effects observed in MOSFETs are next dealt with. Subsequently, the complementary MOS (CMOS) inverter is taken up. Its DC characteristics, noise margin and the small-signal characteristics are discussed. Various load configurations of MOS inverters including passive resistance as well as transistors are presented. The differential inverter involving double-ended inputs and outputs are discussed. The complementary switch or the transmission gate, the tristate inverter and the bipolar devices are briefly dealt with.

2.1.1 nMOS and pMOS Enhancement Transistors

Figure 2.2 depicts a simplified view of the basic structure of an n-channel enhancement mode transistor, which is formed on a p-type substrate of moderate doping level. As shown in the figure, the source and the drain regions made of two isolated islands of n^+-type diffusion. These two diffusion regions are connected via metal to the

external conductors. The depletion regions are mainly formed in the more lightly doped p-region. Thus, the source and the drain are separated from each other by two diodes, as shown in Figure 2.2. A useful device can, however, be made only be maintaining a current between the source and the drain. The region between the two diffused islands under the oxide layer is called the *channel* region. The channel provides a path for the majority carriers (electrons for example, in the n-channel device) to flow between the source and the drain.

The channel is covered by a thin insulating layer of silicon dioxide (SiO_2). The gate electrode, made of polycrystalline silicon (polysilicon or poly in short) stands over this oxide. As the oxide layer is an insulator, the DC current from the gate to the channel is zero. The source and the drain regions are indistinguishable due to the physical symmetry of the structure. The current carriers enter the device through the source terminal while they leave the device by the drain.

The switching behaviour of a MOS device is characterized by an important parameter called the *threshold* voltage (V_{th}), which is defined as the minimum voltage, that must be established between the gate and the source (or between the gate and the substrate, if the source and the substrate are shorted together), to enable the device to conduct (or "turn on"). In the enhancement mode device, the channel is *not* established and the device is in a *non-conducting* (also called *cutoff* or *sub-threshold*) state, for $V_{GS} < V_{th}$. If the gate is connected to a suitable *positive voltage* with respect to the source, then the electric field established between the gate and the source will *induce* a charge inversion region, whereby a conducting path is formed between the source and the drain. In the enhancement mode device, the formation of the channel is *enhanced* in the presence of the gate voltage.

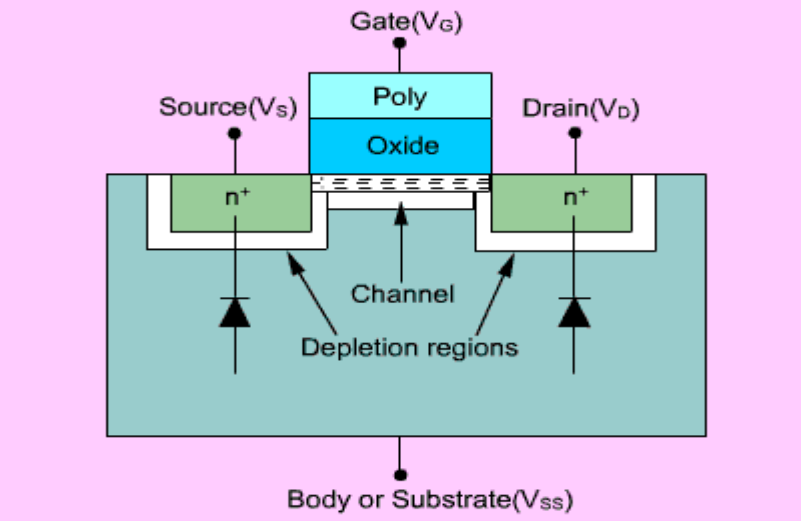


Figure 2.2: Structure of an nMOS enhancement mode transistor. Note that $V_{GS} > V_{th}$, and V_{DS} =0.

By implanting suitable impurities in the region between the source and the drain before depositing the insulating oxide and the gate, a channel can also be established. Thus the source and the drain are connected by a conducting channel even though the voltage between the gate and the source, namely

$V_{GS}=0$ (below the threshold voltage). To make the channel disappear, one has to apply a suitable negative voltage on the gate. As the channel in this device can be *depleted* of the carriers by applying a negative voltage V_{td} say, such a

device is called a *depletion* mode device. Figure 2.3 shows the arrangement in a depletion mode MOS device. For an n-type depletion mode device, penta-valent impurities like phosphorus is used.

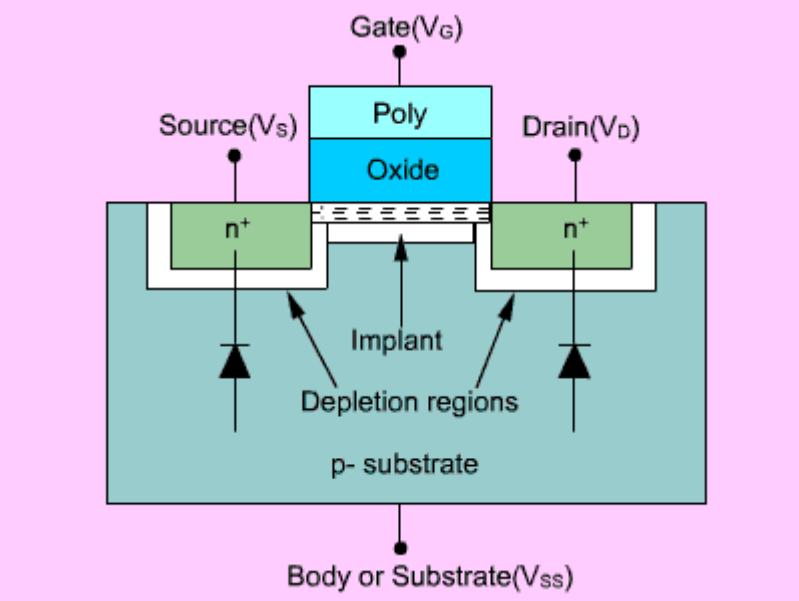


Figure 2.3 Structure of an nMOS depletion mode transistor

To describe the operation of an nMOS enhancement device, note that a positive voltage is applied between the source and the drain (V_{DS}). No current flows from the source and the drain at a zero gate bias (that is, V_{GS}= 0). This is because the source and the drain are insulated from each other by the two reverse-biased diodes as shown in Figure 2.2.However, as a voltage, positive relative to the source and the substrate, is applied to the gate, an electric field is produced across the p-type substrate, This electric field attracts the electrons toward the gate and repels the holes. If the gate voltage is adequately high, the region under the gate changes from p-type to n-type, and it provides a conduction path between the source and the drain. A very thin surface of the p-type substrate is then said to be *inverted,* and the channel is said to be an *n-channel*.

To explain in more detail the electrical behaviour of the MOS structure under external bias, assume that the substrate voltage V_{SS} = 0, and that the gate voltage V_G is the controlling parameter. Three distinct operating regions, namely *accumulation, depletion* and *inversion* are identified based on polarity and magnitude of V_G .

If a negative voltage V_G is applied to the gate electrode, the holes in the p-type substrate are attracted towards the oxide-semiconductor interface. As the majority carrier (hole) concentration near the surface is larger than the equilibrium concentration in the substrate, this condition is referred to as the carrier *accumulation* on the surface. In this case, the oxide electric field is directed towards the gate electrode. Although the hole density increases near the surface in response to the negative gate bias, the minority carrier (electron) concentration goes down as the electrons are repelled deeper into the substrate.

Consider next the situation when a small positive voltage V_G. is applied to the gate. The direction of the electric field across the oxide will now be towards the substrate. The holes (majority carriers) are now driven back into the substrate, leaving the negatively charged immobile acceptor ions. Lack of majority carriers create a *depletion* region near the surface. Almost no mobile carriers are found near the semiconductor-oxide interface under this bias condition.

Next, let us investigate the effect of further increase in the positive gate bias. At a voltage $V_{GS} = V_{th}$, the region near the semiconductor surface acquires the properties of n-type material. This n-type surface layer however, is not due to any

doping operation, but rather by *inversion* of the originally p-type semiconductor owing to the applied voltage. This inverted layer, which is separated from the p-type substrate by a depletion region, accounts for the MOS transistor operation. That is, the thin inversion layer with a large mobile electron concentration, which is brought about by a sufficiently large positive voltage between the gate and the source, can be effectively used for conducting current between the source and the drain terminals of the MOS transistor. *Strong inversion* is said to occur when the concentration of the mobile electrons on the surface equals that of the holes in the underlying p-type substrate.

As far as the electrical characteristics are concerned, an nMOS device acts like a voltage-controlled switch that starts to conduct when V_G (or, the gate voltage with respect to the source) is at least equal to V_{th} (the threshold voltage of the device). Under this condition, with a voltage V_{DS} applied between the source and the drain, the flow of current across the channel occurs as a result of interaction of the electric fields due to the voltages V_{DS} and V_{GS}. The field due to V_{DS} sweeps the electrons from the source toward the drain.As the voltage V_{DS} increases, a resistive drop occurs across the channel. Thus the voltage between the gate and the channel varies with the distance along the channel. This changes the shape of the channel, which becomes tapered towards the drain end.

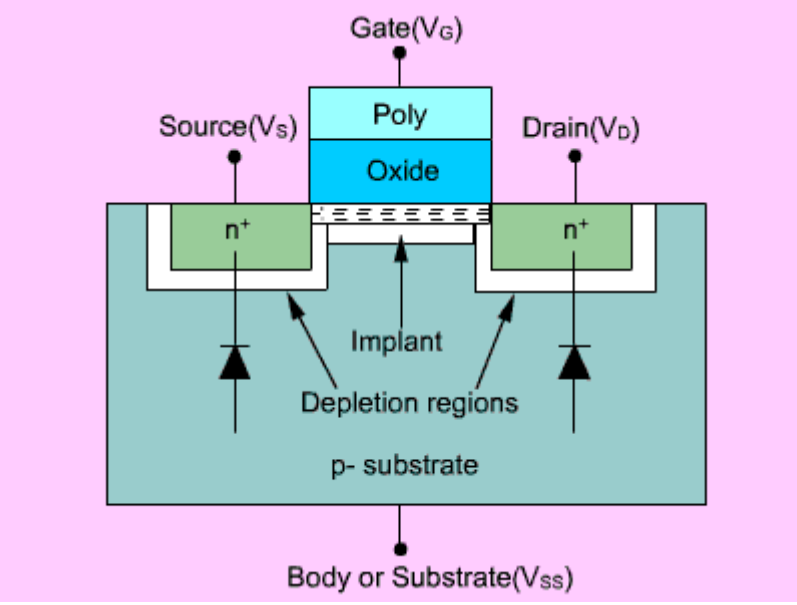


Figure 2.4: An nMOS enhancement mode transistor in non-saturated (linear or resistive) mode. Note that $V_{GS} > V_{th}$, and $V_{DS} < V_{GS} - V_{th}$.

Operating Principles of MOS Transistors

Operating Principles of MOS Transisitors

However, under the circumstance $V_{DS} > V_{GS} - V_{th}$, when the gate voltage relative to drain voltage is insufficient to form the channel (that is, $V_{GD} < V_{th}$), the channel is terminated before the drain end. The channel is then said to be pinched off. This region of operation, known as *saturated* or *pinch-off* condition, is portrayed in Figure 2.5. The effective channel length is thus reduced as the inversion layer near the drain end vanishes. As the majority carriers (electrons) reach the end of the channel, they are swept to the drain by the drift action of the field due to the drain voltage. In the saturated state, the channel current is controlled by the gate voltage and is almost independent of the drain voltage.

In short, the nMOS transistor possesses the three following regions of operation :

- Cutoff, sub-threshold or non-conducting zone
- Non-saturation or linear zone
- Saturation region

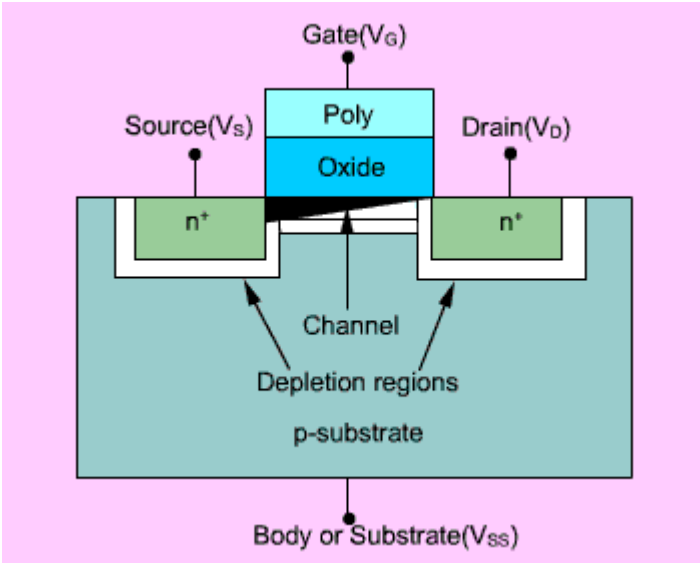


Figure 2.5: An nMOS enhancement mode transistor in saturated (pinch-off) mode. Note that $V_{GS} > V_{th}$, and $V_{DS} > V_{GS} - V_{th}$.

Thus far, we have dealt with principle of operation of an nMOS transistor. A p-channel transistor can be realized by interchanging the n-type and the p-type regions, as shown in Figure 2.6. In case of an pMOS enhancement-mode transistor, the threshold voltage V_{th} is negative. As the gate is made negative with respect to the source by at least $|V_{th}|$, the holes are attracted into the thin region below the gate, crating an inverted *p-channel* . Thus, a conduction path is created for the majority carriers (holes) between the source and the drain. Moreover, a negative drain voltage V DS draws the holes through the channel from the source to the drain.

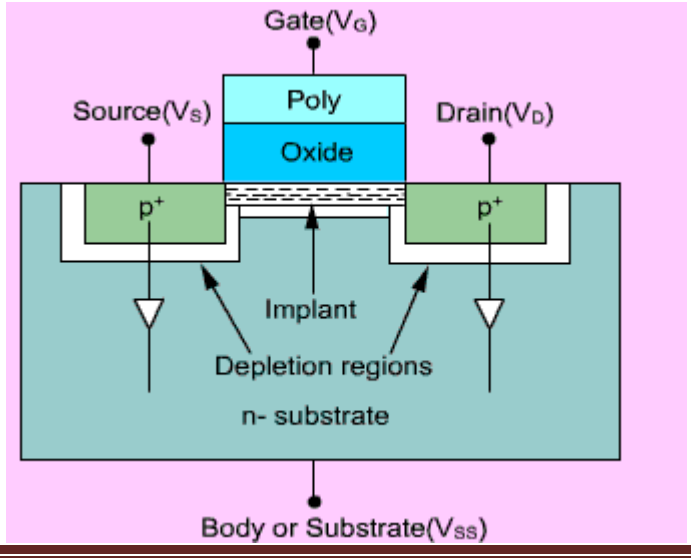


Figure 2.6 Structure of an pMOS enhancement mode transistor. Note that $V_{GS} < V_{th}$, and V_{DS} =0.

2.1.2. Threshold Voltage and Body Effect

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

The physical components determining the threshold voltage are the following.

- work function difference between the gate and the substrate.
- gate voltage portion spent to change the surface potential.
- gate voltage part accounting for the depletion region charge.
- gate voltage component to offset the fixed charges in the gate oxide and the silicon-oxide boundary.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Example 2.1 Given the following parameters, namely the acceptor concentration of p-substrate $N_A = 10^{16}$ cm^{-3}, polysilicon gate doping concentration $N_D = 10^{16}$ cm^{-3}, intrinsic concentration of Si, $n_i = 1.45 \times 10^{10}$ cm^{-3}, gate oxide

thickness t_{ox} =500 Å and oxide-interface fixed charge density N_{ox} =4 X 10^{10}cm^{-2} , calculate the threshold voltage V_{TO} at V_{SB}=0.

$\times 0.35)^{1/2} = -4.82 \times 10^{-8}\ C/cm^2$

Ans:

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

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

For the polysilicon gate, as the doping concentration is extremely high, the heavily doped n-type gate material can be assumed to be degenerate. That is, the Fermi level E_F is almost coincident 7.02×10^{-8} with the bottom of the conduction band E C . Hence, assuming that the intrinsic energy level E_I is at the middle of the band gap, the potential difference between E_I and E_F for the gate is ϕ_F= ½ (energy band gap of Si) = 1/2 X 1.1 = 0.55 V.

Thus, the work function difference ϕ_{GS} 0.40 volt between the doped polysilicon gate and the p-type substrate is -0.35 V - 0.55 V = -0.90 V.

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

$$Q_{B0} = -2\left(\varepsilon_S q N_A \phi_F\right)^{1/2} = -2\left(11.7\times8.85\times10^{-14}\times1.6\times10^{-19}\times10^{16}\right.$$

The oxide-interface charge density is

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

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

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

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

$$V_{TO} = \phi_{GS} - 2\phi_F(\text{substrate}) - Q_{B0}/C_{ox} - Q_i/C_{ox} = -0.90 - (-0.69) - 0.09 =$$

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

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

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

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

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

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

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

$V^{\frac{1}{2}}$

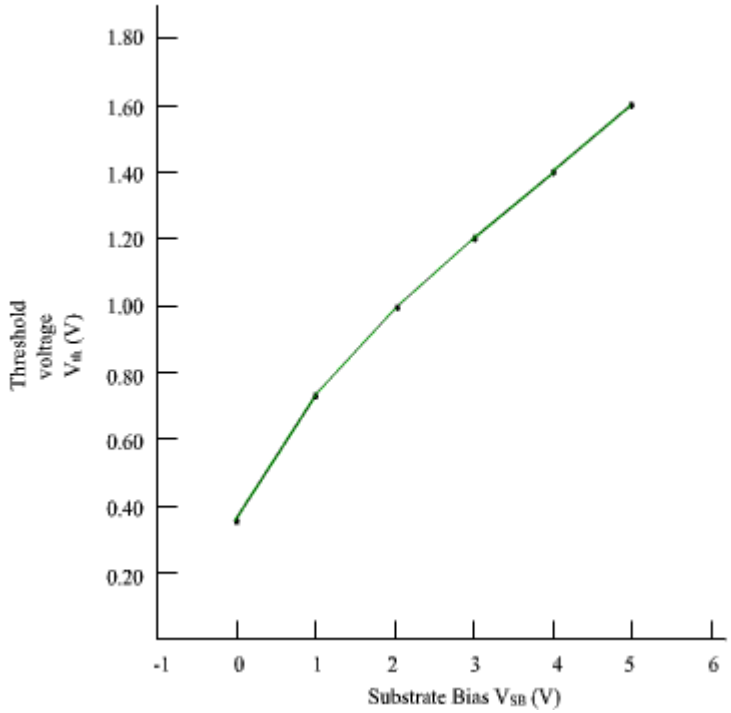


Figure 2.7 Variation of Threshold voltage in response to change in source-to-substrate voltage V_{SB}

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

2.2 MOS Device Current -Voltage Equations

This section first derives the current-voltage relationships for various bias conditions in a MOS transistor. Although the subsequent discussion is centred on an nMOS transistor, the basic expressions can be derived for a pMOS transistor by simply replacing the electron mobility μ_n by the hole mobility μ_p and reversing the polarities of voltages and currents.

As mentioned in the earlier section, the fundamental operation of a MOS transistor arises out of the gate voltage V_{GS} (between the gate and the source) creating a channel between the source and the drain, attracting the majority carriers from the source and causing them to move towards the drain under the influence of an electric field due to the voltage V_{DS} (between the drain and the source). The corresponding current I_{DS} depends on both V_{GS} and V_{DS} .

2.2.1 Basic DC Equations

Let us consider the simplified structure of an nMOS transistor shown in Figure 2.8, in which the majority carriers electrons flow from the source to the drain.

The conventional current flowing from the drain to the source is given by

$I_{DS} = -I_{SD} = (\text{charge induced in channel})/(\text{electron}$

transit

time) $= Q_G / \tau_n$

Now, transit time τ_n = (length of the channel) / (electron velocity) = L / v

where velocity is given by the electron mobility and electric field; or, $v = \mu_n E_{DS}$

Now, $E_{DS} = V_{DS} / L$, so that velocity $v = (\mu_n V_{DS}) / L$

Thus, the transit time is $\tau_n = L^2 / (\mu_n V_{DS})$

At room temperature (300 K), typical values of the electron and hole mobility are given by

$\mu_n = 650 \text{ cm}^2 / V - \text{sec}$, and $\mu_p = 240 \text{ cm}^2 / V - \text{sec}$

We shall derive the current-voltage relationship separately for the linear (or non-saturated) region and the saturated region of operation.

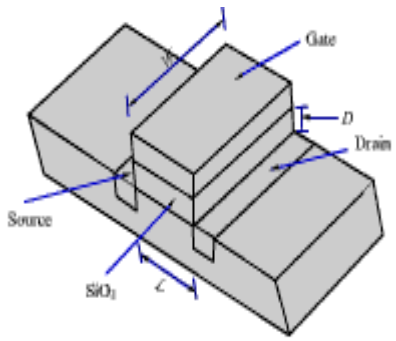


Fig 2.8: Simplified geometrical structure of an nMOS transistor

Linear region : Note that this region of operation implies the existence of the uninterrupted channel between the source and the drain, which is ensured by the voltage relation V_{GS} - V_{th} > V_{DS} .

In the channel, the voltage between the gate and the varies ***linearly*** with the distance x from the source due to the IR drop in the channel. Assume that the device is not saturated and the average channel voltage is V_{DS} /2.

The effective gate voltage $V_{G,eff} = V_{gs}$ - V_{th}

Charge per unit area = $E_g \varepsilon_{ins} \varepsilon_0$

where E_g average electric field from gate to channel, ε_{ins} : relative permittivity of oxide between gate and channel (~4.0

ε_0 $Q_G = E_g \varepsilon_{ins} \varepsilon_0 WL$

for SiO_2), and : free space permittivity (8.85 x 10^{-14} F/cm). So, induced charge .

where W: width of the gate and L : length of channel.

$$Q_C = E_g \varepsilon_{ins} \varepsilon_0 WL$$

$$Q_C = WL\varepsilon_{ins}\varepsilon_0 / D\left\{(V_{GS} - V_{th}) - V_{DS}/2\right\}$$

$$\tau_n = L^2/(\mu_n V_{DS})$$

Thus, the current from the drain to the source may be expressed as

$$I_{DS} = Q_C/\tau_n = \varepsilon_{ins}\varepsilon_0\mu_n W/(LD)\left\{(V_{GS} - V_{th}) - V_{DS}/2\right\}V_{DS}$$

Thus, in the non-saturated region, where $V_{DS} < V_{GS}$

$$I_{DS} = (KW)/L\left\{(V_{GS} - V_{th})V_{DS} - V^2_{DS}/2\right\} \quad \text{.....(2.2)}$$

where the parameter $K = (\varepsilon_{ins}\varepsilon_0\mu_n)/D$

Writing $\beta = (KW)/L$, where W/L is contributed by the geometry of the device,

$$I_{DS} = \beta\left\{(V_{GS} - V_{th})V_{DS} - V^2_{DS}/2\right\} \quad \text{.....(2.3)}$$

Since, the gate-to-channel capacitance is $C_G = (\varepsilon_{ins}\varepsilon_0 WL)/D$ (parallel plate capacitance), then $K = (C_G\mu_n)/(WL)$

, so that (2.2) may be written as

$$I_{DS} = (C_G \mu_n) / L^2 \left\{ (V_{GS} - V_{th}) V_{DS} - V^2_{DS} \right.$$

$/2\}$

.....(2.4)

Denoting $C_G = C_0 WL$ where C_0 : gate capacitance per unit area,

$$I_{DS} = (C_0 \mu_n W)/L^2 \{(V_{GS} - V_{th})V_{DS} - V^2_{DS}/2\} \quad \text{..... (2.5)}$$

Saturated region : Under the voltage condition V_{GS} - V_{th} = V_{DS} , a MOS device is said to be in saturation region of operation. In fact, saturation begins when V_{DS} = V_{GS} - V_{th} , since at this point, the resistive voltage drop (IR drop) in the channel equals the effective gate-to-channel voltage at the drain. One may assume that the current remains *constant* as V_{DS} increases further. Putting V_{DS} = V_{GS} - V_{th} , the equations (2.2-2.5) under saturation condition need to be modified as

$$I_{DS} = (KW)/L\left\{(V_{GS} - V_{th})^2/2\right\} \quad \text{.....(2.6)}$$

$$I_{DS} = \beta(V_{GS} - V_{th})^2/2 \quad \text{.....(2.7)}$$

$$I_{DS} = \left\{C_G \mu_n (V_{GS} - V_{th})^2\right\}/(2L^2) \quad \text{.....(2.8)}$$

$$I_{DS} = \left\{C_0 \mu_n (V_{GS} - V_{th})^2\right\}/(2L) \quad \text{.....(2.9)}$$

The expressions in the last slide derived for I_{DS} are valid for both the enhancement and the depletion mode devices. However, the threshold voltage for the nMOS depletion mode devices (generally denoted as V_{td}) is *negative* .

Figure 2.9 depicts the typical current-voltage characteristics for nMOS enhancement as well as depletion mode transistors. The corresponding curves for a pMOS device may be obtained with appropriate reversal of polarity. For an *n* -channel device with μ_n = 600 cm^2/ V.s, C_0 = 7 X 10^{-8} F/cm^2 , W = 20 m, L = 2 m and V_{th} = V_{T0} = 1.0 V, let us examine the relationship between the drain current and the terminal voltages.

$$K = (C_G \mu_n)/(WL) = (C_0 \mu_n W)/L = 600\,\text{cm}^2/\text{V.s} \times 7 \times 10^{-8}\,\text{F/cm}^2 \times 20\mu\text{m}/2\,\mu\text{m} = 0.42\,\text{mA/V}^2$$

Now, the current-voltage equation (2.2) can be written as follows.

$$I_{DS} = 0.21\,\text{mA/V}^2\left\{2(V_{GS} - 1.0)V_{DS} - V^2_{DS}\right\}$$

If one plots I_{DS} as a function of V_{DS} , for different (constant) values of V_{GS} , one would obtain a characteristic similar to the one shown in Figure 2.9. It may be observed that the second-order current-voltage equation given above gives rise to a set of inverted parabolas for each constant V_{GS} value.

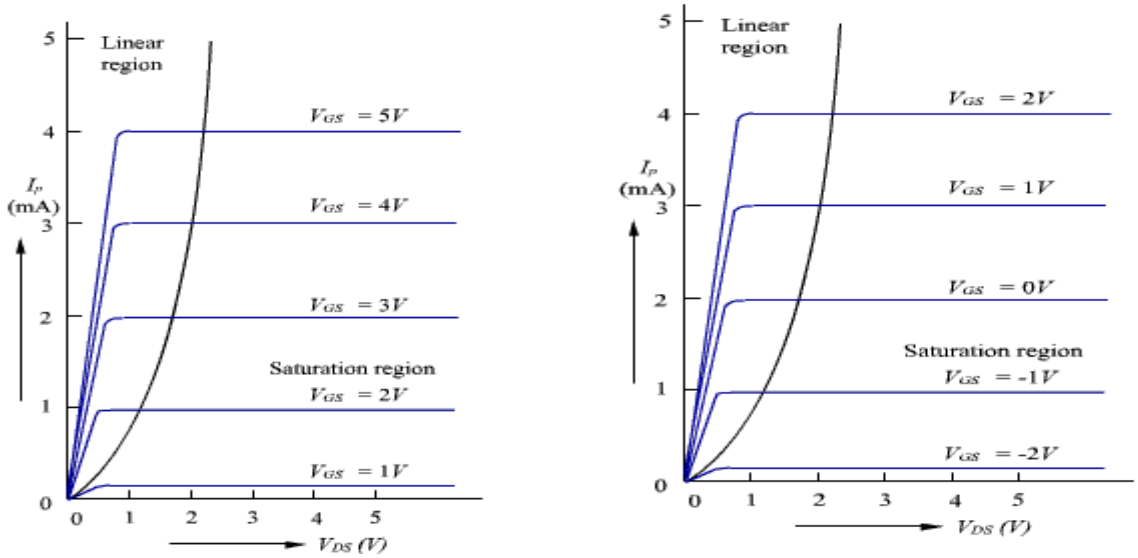
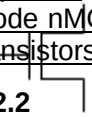


Figure 2.9 Typical current-voltage characteristics for (a) enhancement mode and (b) depletion mode nMOS transistors

Figure in the previous slide:



2.2.2 Second Order Effects

The current-voltage equations in the previous section however are ideal in nature. These have been derived keeping various secondary effects out of consideration.

Threshold voltage and body effect : as has been discussed at length in Sec. 2.1.6, the threshold voltage V_{th} does vary with the voltage difference V_{sb} between the source and the body (substrate). Thus including this difference, the generalized expression for the threshold voltage is reiterated as

$$V_T = V_{T0} + \gamma\left(\sqrt{|2\phi_F + V_{SB}|} - \sqrt{|2\phi_F|}\right) \quad \text{.....} \quad \text{.....} \quad \text{.....} \quad (2.10)$$

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

$$\gamma = \frac{\sqrt{2qN_A\varepsilon_S}}{C_{ox}}$$

$$\frac{t_{ox}}{\varepsilon_{ox}}\sqrt{2qN_A\varepsilon_S} = \frac{\gamma}{C_{ox}}\sqrt{2qN_A\varepsilon_S}$$

$11.7 \times 8.85 \times 10^{-14}$ 1.6×10^{-9}

.Typical values of range from 0.4 to 1.2. It may also be written as

Example 2.3: $N_A = 3\times10^{16}\ \text{cm}^{-3},\ t_{ox} = 200\ \overset{0}{A},\ \varepsilon_{ox} =$ and

$$\gamma = \frac{0.2\times10^{-5}}{3.9\times8.85\times10^{-14}}\sqrt{2\times16\times10^{-19}\times11.7\times8.85\times10^{-14}\times3\times10^{16}} =$$

$$\phi_b = 0.0261\ln\left(\frac{3\times10^{16}}{1.5\times10^{10}}\right) = 0.375$$

Then, at V_{sb} = 2.5 volts

$$V_{T2.5} = V_{T0} + 0.57\left(\sqrt{0.75+2.5} - \sqrt{0.75}\right) = V_{T0} + 0.53$$

As is clear, the threshold voltage increases by almost half a volt for the above process parameters when the source is higher than the substrate by 2.5 volts.

Drain punch-through : In a MOSFET device with improperly scaled small channel length and too low channel doping, undesired electrostatic interaction can take place between the source and the drain known as *drain-induced barrier lowering* (DIBL) takes place. This leads to punch-through leakage or breakdown between the source and the drain, and

loss of gate control. One should consider the surface potential along the channel to understand the punch-through phenomenon. As the drain bias increases, the conduction band edge (which represents the electron energies) in the drain is pulled down, leading to an increase in the drain-channel depletion width.

In a long-channel device, the drain bias does not influence the source-to-channel potential barrier, and it depends on the increase of gate bias to cause the drain current to flow. However, in a short-channel device, as a result of increase in drain bias and pull-down of the conduction band edge, the source-channel potential barrier is lowered due to DIBL. This in turn causes drain current to flow regardless of the gate voltage (that is, even if it is below the threshold voltage V_{th}). More simply, the advent of DIBL may be explained by the expansion of drain depletion region and its eventual merging with source depletion region, causing punch-through breakdown between the source and the drain. The punch-through condition puts a natural constraint on the voltages across the internal circuit nodes.

Sub-threshold region conduction : the cutoff region of operation is also referred to as the sub-threshold region, which is mathematically expressed as $I_{DS} = 0 \qquad V_{GS} < V_{th}$

However, a phenomenon called *sub-threshold conduction* is observed in small-geometry transistors. The current flow in the channel depends on creating and maintaining an inversion layer on the surface. If the gate voltage is inadequate to invert the surface (that is, $V_{GS} < V_{T0}$), the electrons in the channel encounter a *potential barrier* that blocks the flow. However, in small-geometry MOSFETs, this potential barrier is controlled by both V_{GS} and V_{DS} . If the drain voltage is increased, the potential barrier in the channel decreases, leading to *drain-induced barrier lowering* (DIBL). The lowered potential barrier finally leads to flow of electrons between the source and the drain, even if $V_{GS} < V_{T0}$ (that is, even when the surface is not in strong inversion). The channel current flowing in this condition is called the *sub-threshold current* . This current, due mainly to diffusion between the source and the drain, is causing concern in deep sub-micron designs. The model implemented in SPICE brings in an exponential, semi-empirical dependence of the drain current on V_{GS} in the *weak inversion region.* Defining a voltage V on as the boundary between the regions of weak and strong inversion,

$$I_D(weak\ inversion) = I_{on} \cdot e^{(V_{GS}-V_{on})\left(\frac{q}{nkT}\right)}$$

where I_{on} is the current in strong inversion for $V_{GS} = V_{on}$.

Channel length modulation : so far one has not considered the variations in channel length due to the changes in drain-to-source voltage V_{DS} . For long-channel transistors, the effect of channel length variation is not prominent. With the decrease in channel lenghth, however, the variation matters. Figure 2.5 shows that the inversion layer reduces to a point at the drain end when $V_{DS} = V_{DSAT} = V_{GS} - V_{th}$. That is, the channel is *pinched off* at the drain end. The onset of saturation mode operation is indicated by the pinch-off event. If the drain-to-source voltage is increased beyond the saturation edge ($V_{DS} > V_{DSAT}$), a still larger portion of the channel becomes pinched off. Let the effective channel (that is, the length of the inversion layer) be $L_{eff} = L - \Delta L$.

where L : original channel length (the device being in non-saturated mode), and ΔL : length of the channel segment where the inversion layer charge is zero. Thus, the pinch-off point moves from the drain end toward V_{DS} the source with increasing drain-to-source voltage . The remaining portion of the channel between the pinch-off point and the drain end will be in depletion mode. For the shortened channel, with an effective channel voltage of V_{DSAT} , the channel current is given by

$$I_{DS(SAT)} = \frac{\mu_n C_{ox}}{2} \cdot \frac{W}{L_{eff}} \cdot (V_{GS} - V_{T0})^2 \quad \text{..... (2.11)}$$

The current expression pertains to a MOSFET with effective channel length L_{eff}, operating in saturation. The above equation depicts the condition known as *channel length modulation* , where the channel is reduced in length. As the effective length decreases with increasing V_{DS} , the saturation current $I_{DS(SAT)}$ will consequently increase with increasing V_{DS} . The current given by (2.11) can be re-written as

$$I_{DS(SAT)} = \frac{\mu_n C_{ox}}{2} \cdot \left(\frac{1}{1 - \frac{\Delta L}{L}}\right) \frac{W}{L} \cdot (V_{GS} - V_{T0})^2 \quad \text{..... (2.12)}$$

The second term on the right hand side of (2.12) accounts for the channel modulation effect. It can be shown that the factor channel length ΔL is expressible as

$$\Delta L \propto \sqrt{V_{DS} - V_{DSAT}}$$

One can even use the empirical relation between ΔL and V_{DS} given as follows.

$$1 - \frac{\Delta L}{L} \approx 1 - \lambda V_{DS}$$

The parameter λ is called the *channel length modulation coefficient,* having a value in the range 0.02V -1 to 0.005V -1 .

Assuming that $\lambda V_{DS} << 1$, the saturation current given in (2.11) can be written as

$$I_{DS(SAT)} = \frac{\mu_n C_{ox}}{2} \cdot \frac{W}{L_{eff}} \cdot (V_{GS} - V_{T0})^2 \cdot (1 + \lambda V_{DS})$$

The simplified equation (2.13) points to a linear dependence of the saturation current on the drain-to-source voltage. The slope of the current-voltage characteristic in the saturation region is determined by the channel length modulation factor λ.

Impact ionization :An electron traveling from the source to the drain along the channel gains kinetic energy at the cost of electrostatic potential energy in the pinch-off region, and becomes a “hot” electron. As the hot electrons travel towards the drain, they can create secondary electron-hole pairs by impact ionization. The secondary electrons are collected at the drain, and cause the drain current in saturation to increase with drain bias at high voltages, thus leading to a fall in the output impedance. The secondary holes are collected as substrate current. This effect is called *impact ionization* . The hot electrons can even penetrate the gate oxide, causing a gate current. This finally leads to degradation in MOSFET parameters like increase of threshold voltage and decrease of transconductance. Impact ionization can create circuit problems such as noise in mixed-signal systems, poor refresh times in dynamic memories, or latch-up in CMOS circuits. The remedy to this problem is to use a device with lightly doped drain. By reducing the doping density in the source/drain, the depletion width at the reverse-biased drain-channel junction is increase and consequently, the electric filed is reduced. Hot carrier effects do not normally present an acute problem for *p* -channel MOSFETs. This is because the channel mobility of holes is almost half that of the electrons. Thus, for the same filed, there are fewer hot holes than hot electrons. However, lower hole mobility results in lower drive currents in *p* -channel devices than in *n* -channel devices.

Complementary CMOS Inverter - DC Characteristics

A complementary CMOS inverter is implemented as the series connection of a *p*-device and an *n*-device, as shown in Figure 2.10. Note that the source and the substrate (body) of the *p* -device is tied to the V_{DD} rail, while the source and the substrate of the *n*-device are connected to the ground bus. Thus, the devices do not suffer from any body effect. To derive the DC transfer characteristics for the CMOS inverter, which depicts the variation of the output voltage (V_{out}) as a function of the input voltage (V_{in}), one can identify five following regions of operation for the *n* -transistor and *p* -transistor.

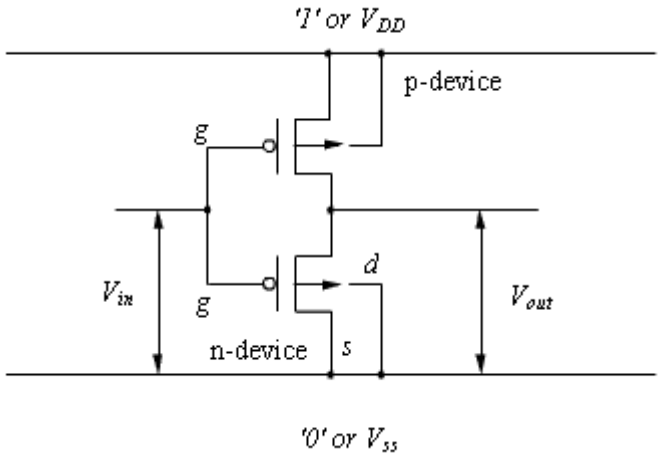


Figure 2.10 A CMOS inverter shown with substrate Connections

Let V_{tn} and V_{tp} denote the threshold voltages of the *n* and *p*-devices respectively. The following voltages at the gate and the drain of the two devices (relative to their respective sources) are all referred with respect to the ground (or V_{SS}), which is the substrate voltage of the *n* -device, namely

$V_{gsn} = V_{in}$, $V_{dsn} = V_{out}$, $V_{gsp} = V_{in} - V_{DD}$, and $V_{dsp} = V_{out} - V_{DD}$.

The voltage transfer characteristic of the CMOS inverter is now derived with reference to the following five regions of operation :

Region 1 : the input voltage is in the range $0 \le V_{in} < V_{tn}$. In this condition, the *n* -transistor is off, while the *p* -transistor is in linear region (as $-V_{DD} < V_{gsp} < -V_{DD} + V_{tn}$).

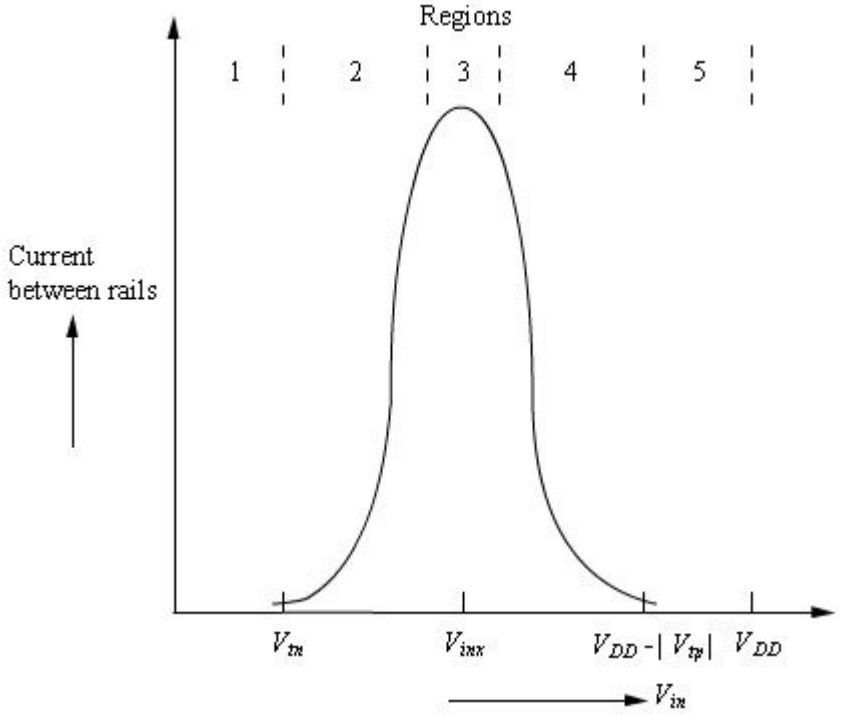


Figure 2.11: Variation of current in CMOS inverter with V_{in}

No actual current flows until V_{in} crosses V_{tn} , as may be seen from Figure 2.11. The operating point of the *p* -transistor moves from higher to lower values of currents in linear zone. The output voltage is given by $V_{out} \approx V_{DD}$, as may be seen from Figure 2.12.

Region 2 : the input voltage is in the range $V_{tn} \le V_{in} < V_{inv}$. The upper limit of V_{in} is V_{inv} , the *logic threshold voltage* of the inverter. The logic threshold voltage or the *switching point voltage* of an inverter denotes the boundary of "logic 1" and "logic 0". It is the output voltage at which $V_{in} = V_{out}$. In this region, the *n*-transistor moves into saturation, while the *p*-transistor remains in linear region. The total current through the inverter increases, and the output voltage tends to drop fast.

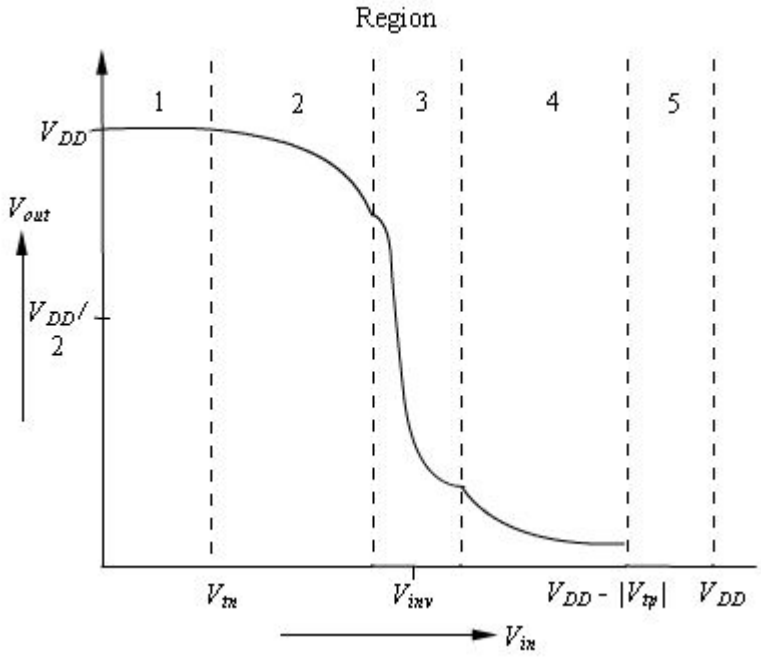


Figure 2.12 Transfer characteristics of the CMOS inverter

Region 3 : In this region, $V_{in} \approx V_{inv}$. Both the transistors are in saturation, the drain current attains a maximum value, and the output voltage falls rapidly. The inverter exhibits gain. But this region is inherently unstable. As both the transistors are in saturation, equating their currents, one gets (as $V_{gsn} = V_{inv}$, $V_{gsp} = V_{inv} - V_{DD}$).

$$\frac{1}{2}\beta_n (V_{inv} - V_{tn})^2 = \frac{1}{2}\beta_p \left[(V_{inv} - V_{DD} - V_{tp}\right)^2 \quad \text{.....(2.14)}$$

where $\beta = K\frac{W}{L}$ and $K = \frac{\varepsilon_{ins}\varepsilon_0\mu}{D}$. Solving for the logic threshold voltage V_{inv}, one gets

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

Note that if $\beta_n = \beta_p$ and $V_{tn} = -V_{tp}$, then V_{inv} =0.5 V_{DD} .

2.15)

Region 4 : In this region, $V_{inv} < V_{in} \leq V_{DD} \quad |V_{tp}|$. As the input voltage V_{in} is increased beyond V_{inv} , the *n* -transistor leaves saturation region and enters linear region, while the *p* -transistor continues in saturation. The magnitude of both the drain current and the output voltage drops.

Region 5 : In this region, $V_{DD} - |V_{tp}| \leq V_{in} \leq V_{DD}$. At this point, the *p* -transistor is turned off, and the *n* -transistor is in linear region, drawing a small current, which falls to zero as V_{in} increases beyond V_{DD} -| V_{tp}|, since the *p* -transistor turns off the current path. The output in this region is $V_{out} \approx 0$.

As may be seen from the transfer curve in Figure 2.12, the transition from "logic 1" state (represented by regions 1 and 2) to “logic 0” state (represented by regions 4 and 5) is quite steep. This characteristic guarantees maximum noise immunity.

β_n / β_p **ratio** : One can explore the variation of the transfer characteristic as a function of the ratio β_n / β_p. As noted from (2.15), the logic threshold voltage V_{inv} depends on the ratio β_n / β_p. The CMOS inverter with the ratio β_n / β_p=1 allows a capacitive load to charge and discharge in equal times by providing equal current-source and current-sink capabilities. Consider the case of β_n / β_p>1. Keeping β_p fixed, if one increases β_n, then the impedance of the pull-down *n* -transistor decreases. It conducts faster, leading to faster discharge of the capacitive load. This ensures quicker fall of the output voltage V_{out} , as V_{in} increases from 0 volt onwards. That is, the transfer characteristic shifts leftwards. Similarly, for a CMOS inverter with β_n / β_p<1, the transfer curve shifts rightwards. This is portrayed in Figure 2.13.

Noise margin : is a parameter intimately related to the transfer characteristics. It allows one to estimate the allowable noise voltage on the input of a gate so that the output will not be affected. Noise margin (also called noise immunity) is specified in terms of two parameters - the low noise margin NM_L , and the high noise margin NM_H . Referring to Figure 2.14, NM_l is defined as the difference in magnitude between the maximum LOW input voltage recognized by the driven gate and the maximum LOW output voltage of the driving gate. That is,

$$NM_L = | V_{ILmax} - V_{OLmax} |$$

Similarly, the value of NM_H is the difference in magnitude between the minimum HIGH output voltage of the driving gate and the minimum HIGH input voltage recognizable by the driven gate. That is,

$$NM_H = | V_{OHmin} - V_{IHmin} |$$

Where V_{IHmin} : minimum HIGH input voltage

V_{ILmax} : maximum LOW input voltage

V_{OHmin} : minimum HIGH output voltage

V_{OLmax} : maximum LOW output voltage

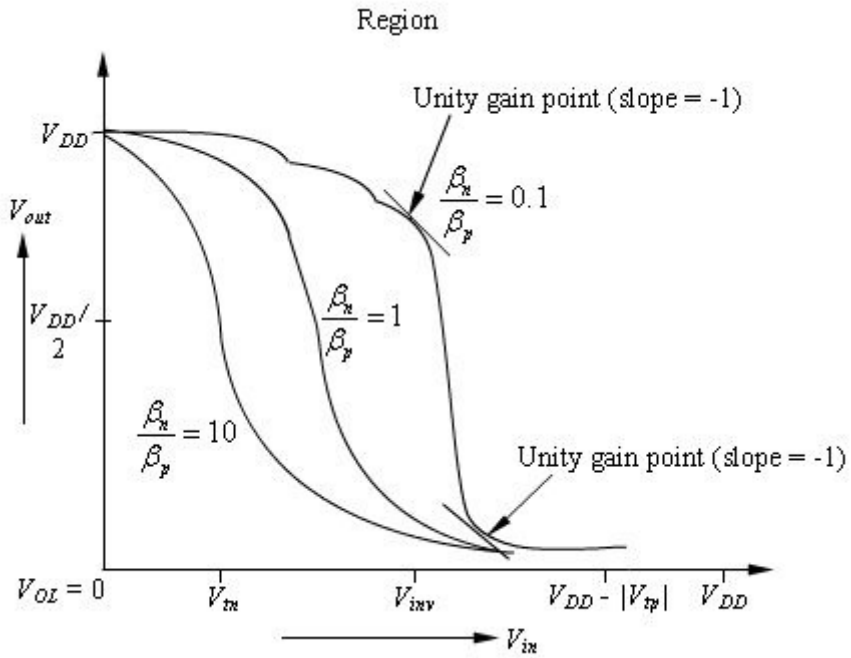


Figure 2.13 Variation of shape of transfer characteristic of the CMOS inverter with the ratio β_n / β_p

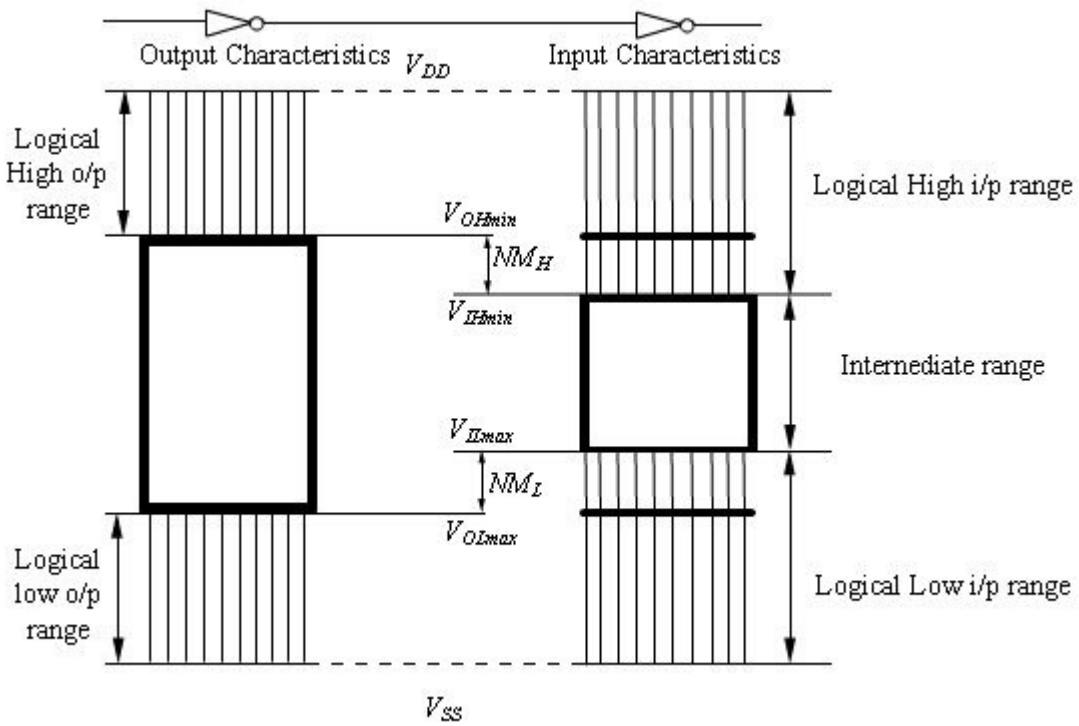


Figure 2.14 Definition of noise margin

Amplifiers with Active Loads – CMOS Amplifiers

Section 3.1 Amplifiers with Active Loads

In the last chapter, we noticed that the load R_L
here: (1) For IC design, this is not desirable because it is not cost effective to fabricate a desired resistor, not mentioning a large resistor will require a rather large space in the IC. (2) A large resistor may easily drive the transistor out of saturation as shown in Fig. 3.1-1.

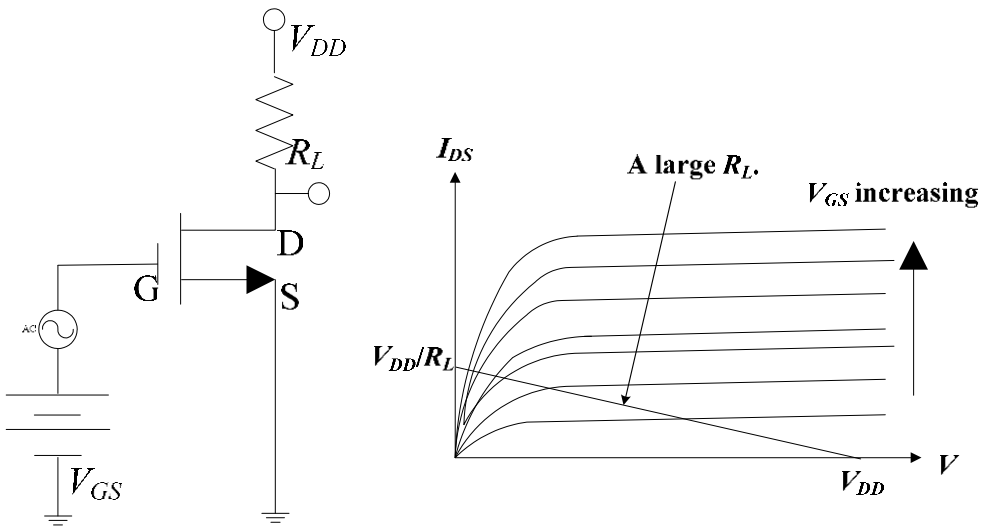


Fig. 3.1-1 A large

It will be desirable if we have a load curve, instead of a load line as shown in Fig. 3.1-2 below:

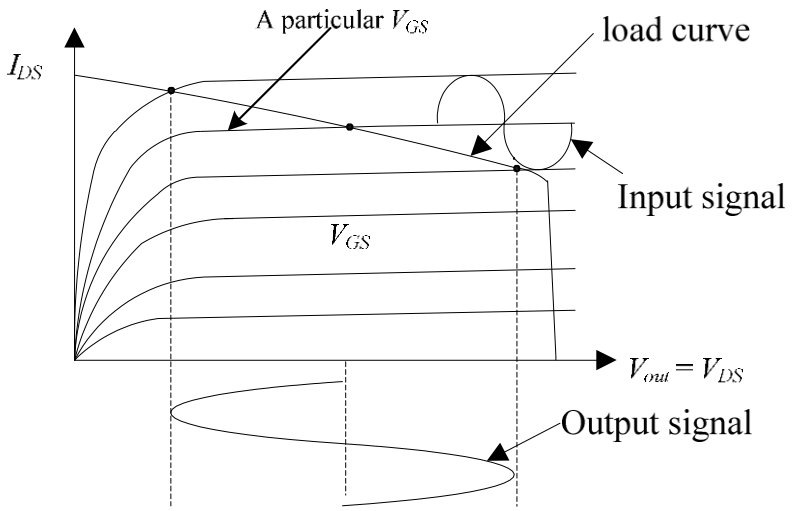


Fig. 3.1-2 A desirable load curve

To achieve this desirable load curve, we may use an active load, instead of a passive load, such as a resistor.

Let us consider the following PMOS and its I-V curve as shown in Fig. 3.1-3.

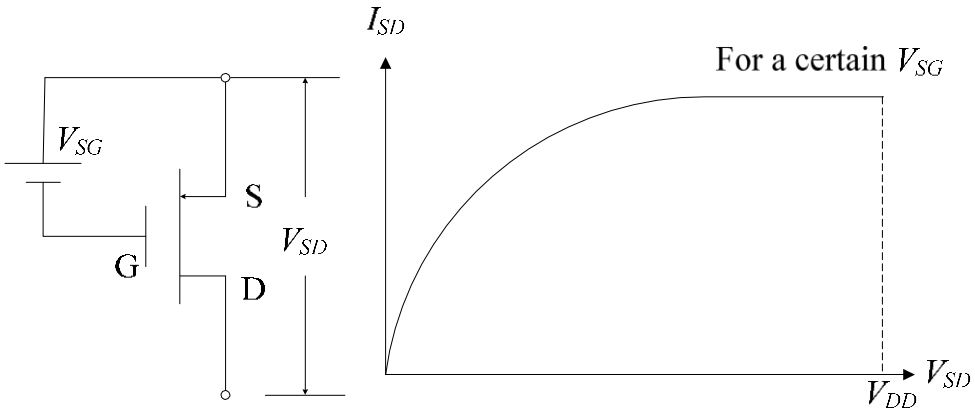


Fig. 3.1-3 A PMOS transistor and its I-V curve

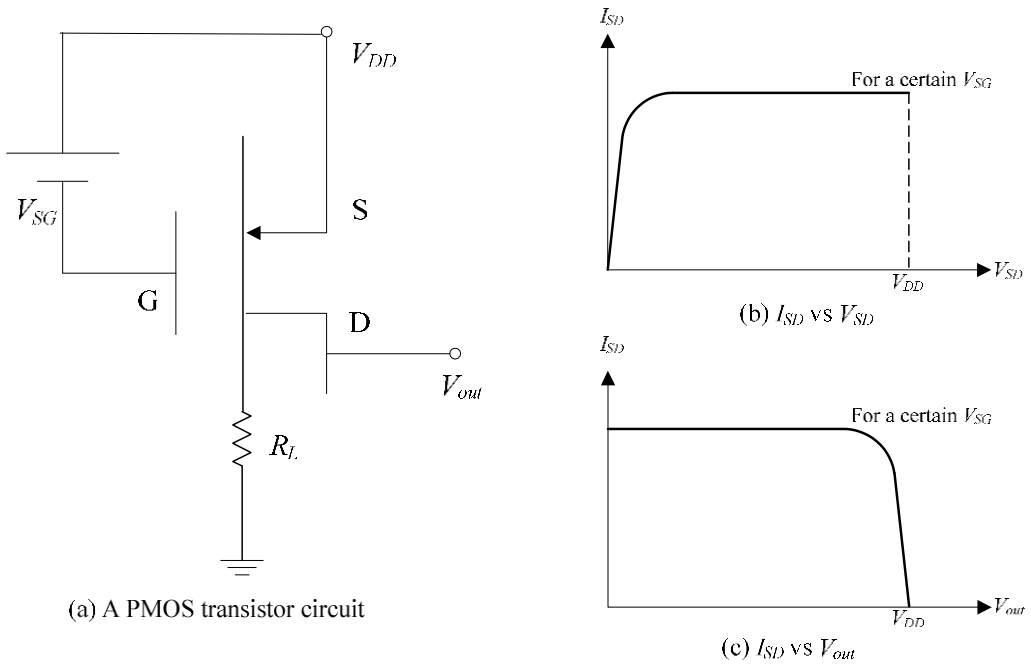


Fig. 3.1-4 A PMOS transistor circuit with its I-V diagrams

From Fig. 3.1-4, we can see that a PMOS circuit can be used as a load for an NMOS amplifier, as shown in Fig. 3.1-5.

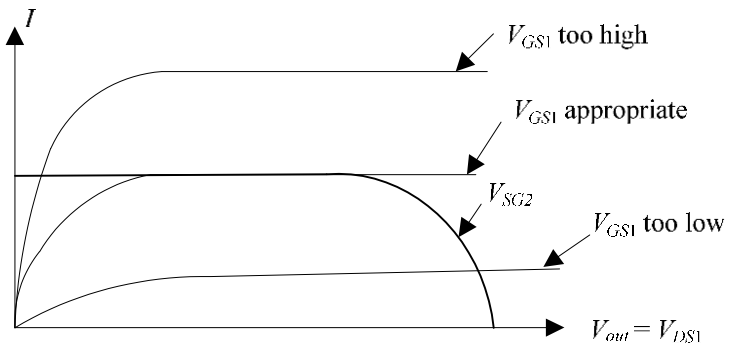
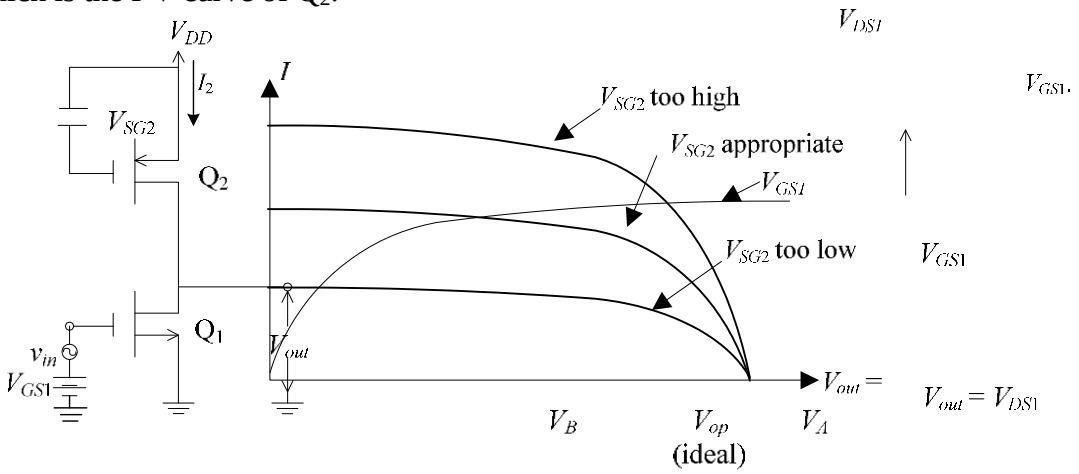


Fig. 3.1-6 Different V_{GS1}'s for a fixed V_{SG2}

Fig. 3.1-7 Different V_{SG2}'s for a fixed V_{GS1}

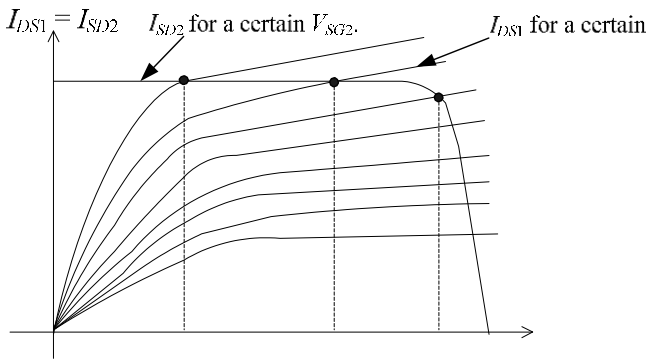
Note that so far as Q_1 is concerned, Q_2 is its load and vice versa, as shown in the above figures. Since NMOS and PMOS are complementary to each other, we call this kind of circuits CMOS circuits.

For the CMOS amplifier shown in Fig. 3.1-5, let us assume that the circuit is properly biased. Fig. 3.1-8 shows the diagram of the I-V curves of Q_1 and its load curve, which is the I-V curve of Q_2.



(a) A CMOS transistor circuit (b) I-V curves of Q_1 and its load curve

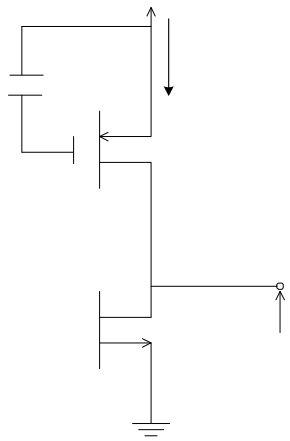
.

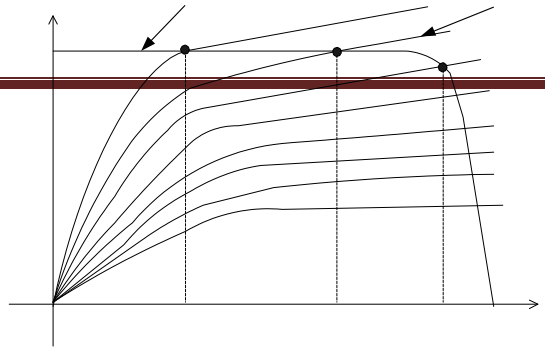


$\frac{}{2}$ A B

behaves as an amplifier..

V_{DI}





I_2

V_{SG2}

Q

V_B V_{op} V_A
(ideal)

$V_{out} = V_{DS1}$

$V_{out} = V_{DS}$

b) I-V curves of Q_1 and the load curve of Q_1

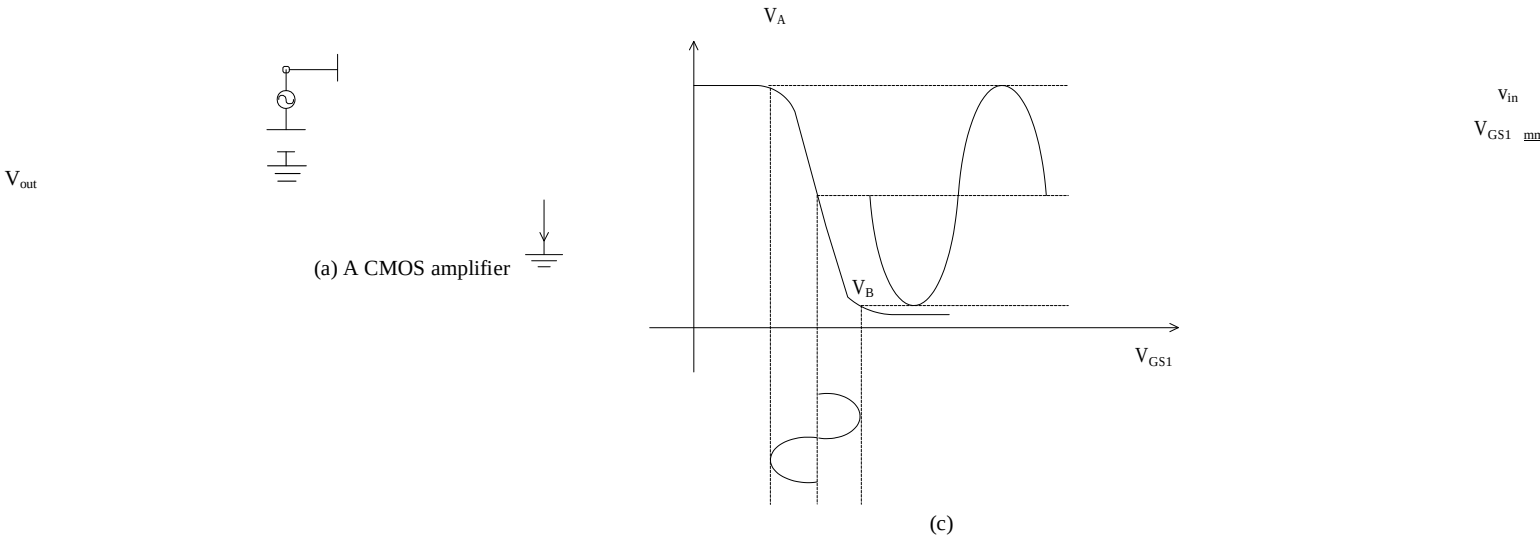
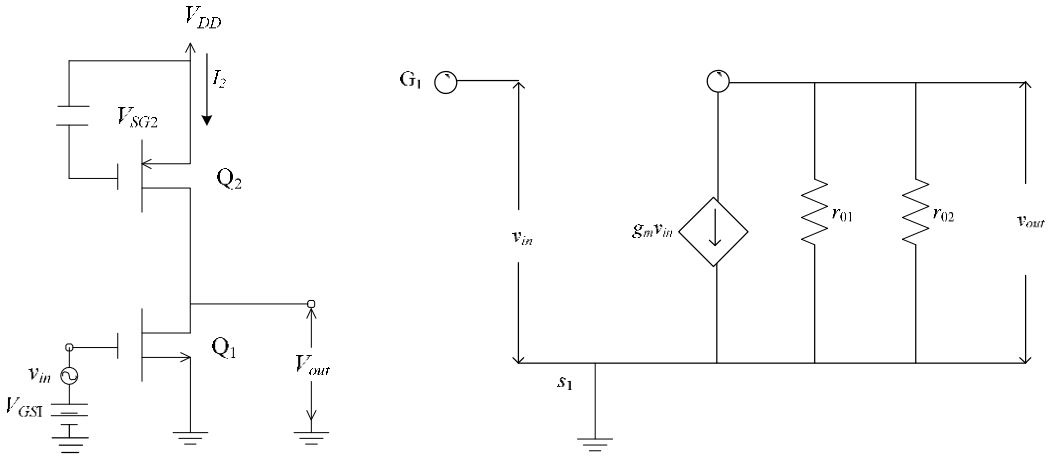


Fig. 3.1-9 The amplification of input signal

The small signal equivalent circuit of the CMOS amplifier is shown in Fig. 3.1-

10. The impedances

For r_{o1} and r_{o2}, refer to Section 2.4.



(a) A CMOS transistor circuit (b) The small signal equivalent circuit of the CMOS amplifier

Fig. 3.1-10 A CMOS transistor circuit and its small signal equivalent circuit

As can be seen,

$$v_{out} = -g_m v_{in}(r_{01} \,//\, r_{02}) \qquad (3.1\text{-}1)$$

If $r_{01} \approx r_{02}$, which is often the case, we have

$$A_V = \frac{v_{out}}{v_{in}} = -\frac{1}{2} g_m r_{01} \qquad (3.1\text{-}2)$$

If a passive load is used, $A_V = g_m R_L$. Since r_{01} is much larger than R_L which can be used, we have obtained a larger gain. By passive loads, we mean loads such as resistors, inductors and capacitors which do not require power supplies.

Section 3.2 Some Experiments about CMOS Amplifiers

The following circuit shown in Fig. 3.2-1 will be used in our SPICE simulation experiments.

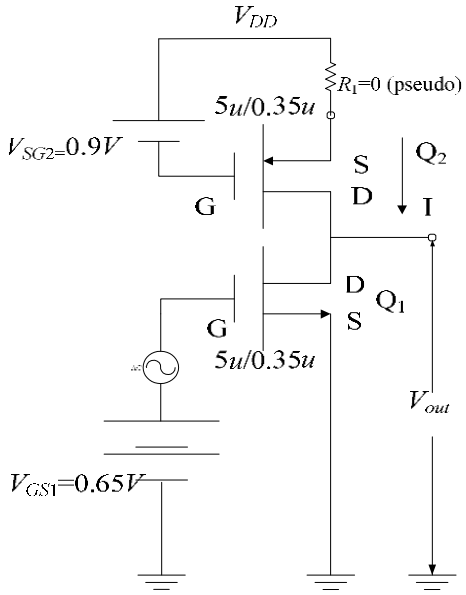


Fig. 3.2-1 The CMOS amplifier circuit for the Experiments in Section 3.2

Experiment 3.2-1. **The I-V Curve of Q_1 and the its Load Curve**.

In Table 3.2-1, we display the SPICE simulation program of the experiment and in Fig. 3.2-2, we show the I-V curve of Q_1 and its load curve. Note that the load curve of Q_1 is the I-V curve of Q_2.

Table 3.2-1 Program of Experiment 3.2-1

```
simple
.protect
.lib 'c:\mm0355v.l' TT
.unprotect
.op
.options nomod post
VDD   11      0       3.3v
R1    111       0k VSG2
      11        2     0.9v
V3    3       0       0v
.param W1=5u
M1    3       4       0       0
+nch L=0.35u W='W1' m=1
+AD='0.95u*W1' PD='2*(0.95u+W1)'
+AS='0.95u*W1' PS='2*(0.95u+W1)'
M2    3       2       1       1
+pch L=0.35u W='W1' m=1
+AD='0.95u*W1' PD='2*(0.95u+W1)'
+AS='0.95u*W1' PS='2*(0.95u+W1)'
```

```
VGS1  4        0        0.65v
.DC V3 0 3.3v 0.1v
```

```
.PROBE I(M2) I(M1) I(R1)
.end
```

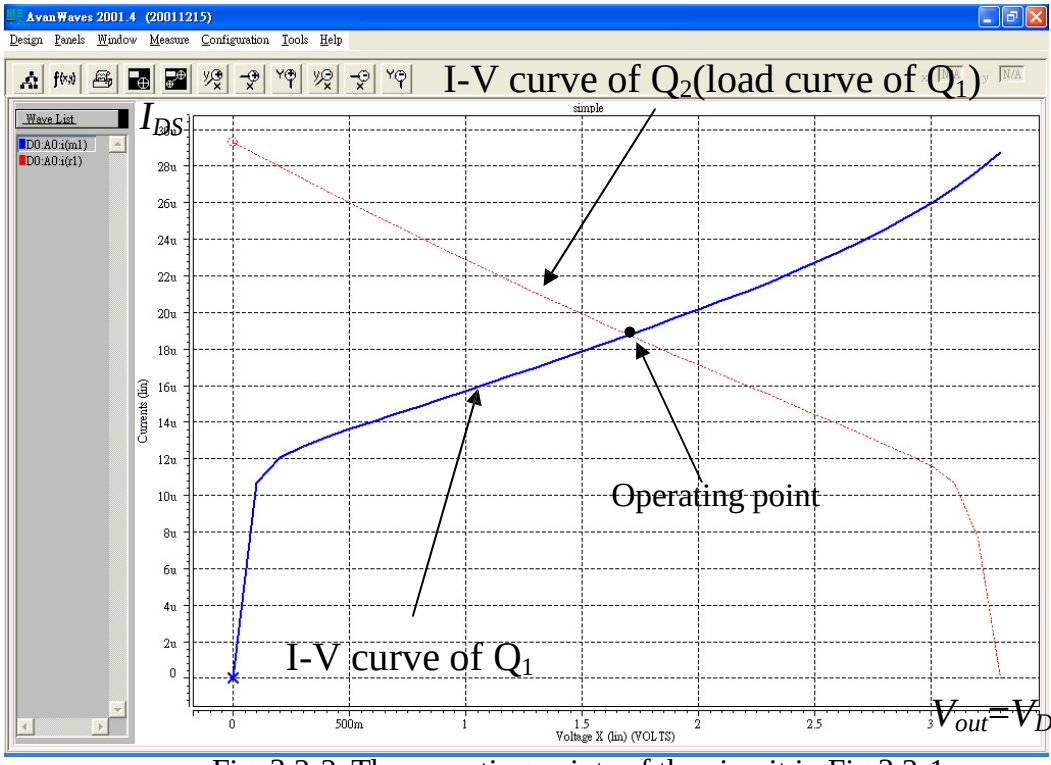


Fig. 3.2-2 The operating points of the circuit in Fig 3.2-1

Experiment 3.2-2 The Operating Point with the Same V_{GS1} and a Smaller V_{SG2}.

In this experiment, we lowered V_{SG2} from 0.9V to 0.8V. The program is shown in Table 3.2-2 and the resulting operating point can be seen in Fig. 3.2-3. In fact, this operating point is close to the ohmic region, which is undesirable.

Table 3.2-2 Program of Experiment 3.2-2

```
simple
.protect
.lib 'c:\mm0355v.l' TT
.unprotect
.op
.options nomod post
VDD   11      0       3.3v
R1    111       0k VSG2
      11        2     0.8v
V3    3       0       0v
.param W1=5u
M1    3       4       0       0
```

```
+nch L=0.35u W='W1' m=1
```

```
+AD='0.95u*W1' PD='2*(0.95u+W1)'
+AS='0.95u*W1' PS='2*(0.95u+W1)'
M2      3         2         1         1
+pch L=0.35u  W='W1' m=1
+AD='0.95u*W1' PD='2*(0.95u+W1)'
+AS='0.95u*W1' PS='2*(0.95u+W1)'
VGS1  4         0         0.65v
.DC V3 0 3.3v 0.1v
.PROBE I(M2) I(M1) I(R1)
.end
```

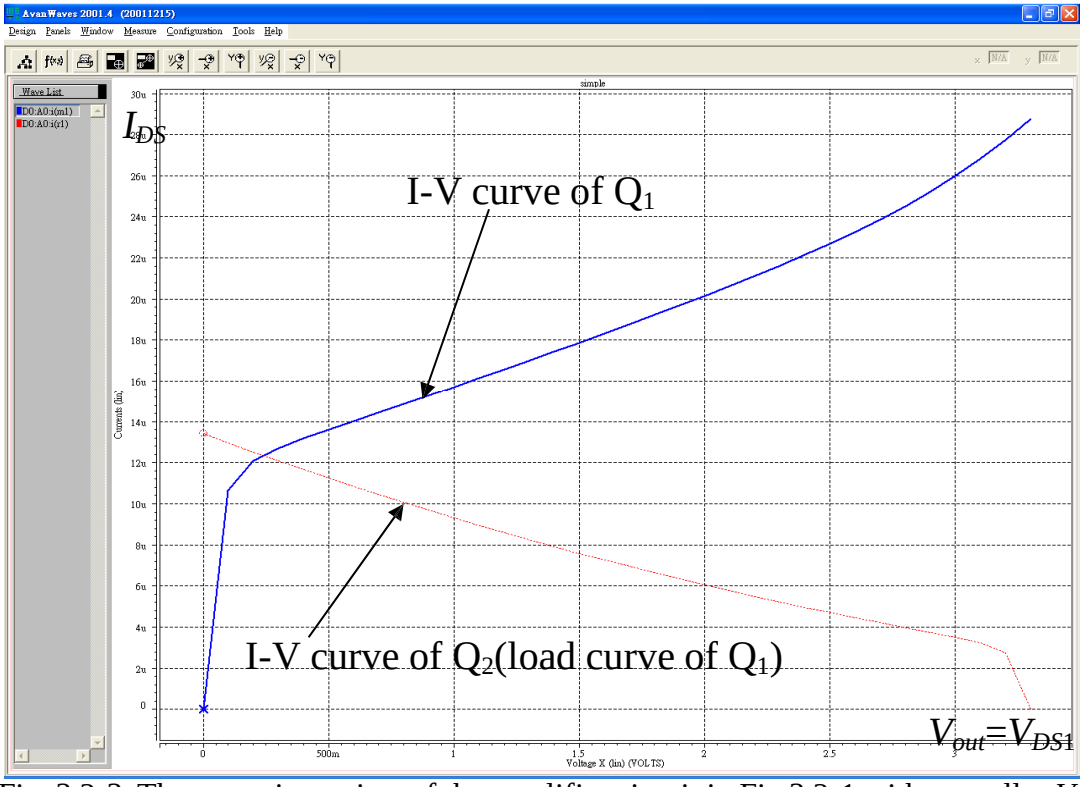


Fig. 3.2-3 The operating points of the amplifier circuit in Fig 3.2-1 with a smaller $V_{SG\,2}$

Experiment 3.2-3 The Operating Point with the Same V_{GS1} and a Higher V_{SG2}

In this experiment, we increased $V_{SG\,2}$ from 0.9V to 1.0V. The program is displayed in Table 3.2-3 and the result is in Fig. 3.2-4. Again, as can be seen, this new operating point is not ideal either.

.unprotect

```
simple
.protect
.lib 'c:\mm0355v.l' TT
```

Table 3.2-3 Program of Experiment 3.2-3

```
.op
.options nomod post
VDD   11      0       3.3v
R1    111       0k VSG2
      11        2     1v
V3    3       0       0v
.param W1=5u
M1    3       4       0       0
+nch L=0.35u W='W1' m=1
+AD='0.95u*W1' PD='2*(0.95u+W1)'
+AS='0.95u*W1' PS='2*(0.95u+W1)'
M2    3       2       1       1
+pch L=0.35u W='W1' m=1
+AD='0.95u*W1' PD='2*(0.95u+W1)'
+AS='0.95u*W1' PS='2*(0.95u+W1)'
VGS1  4       0       0.65v
.DC V3 0 3.3v 0.1v
.PROBE I(M2) I(M1) I(R1)
.end
```

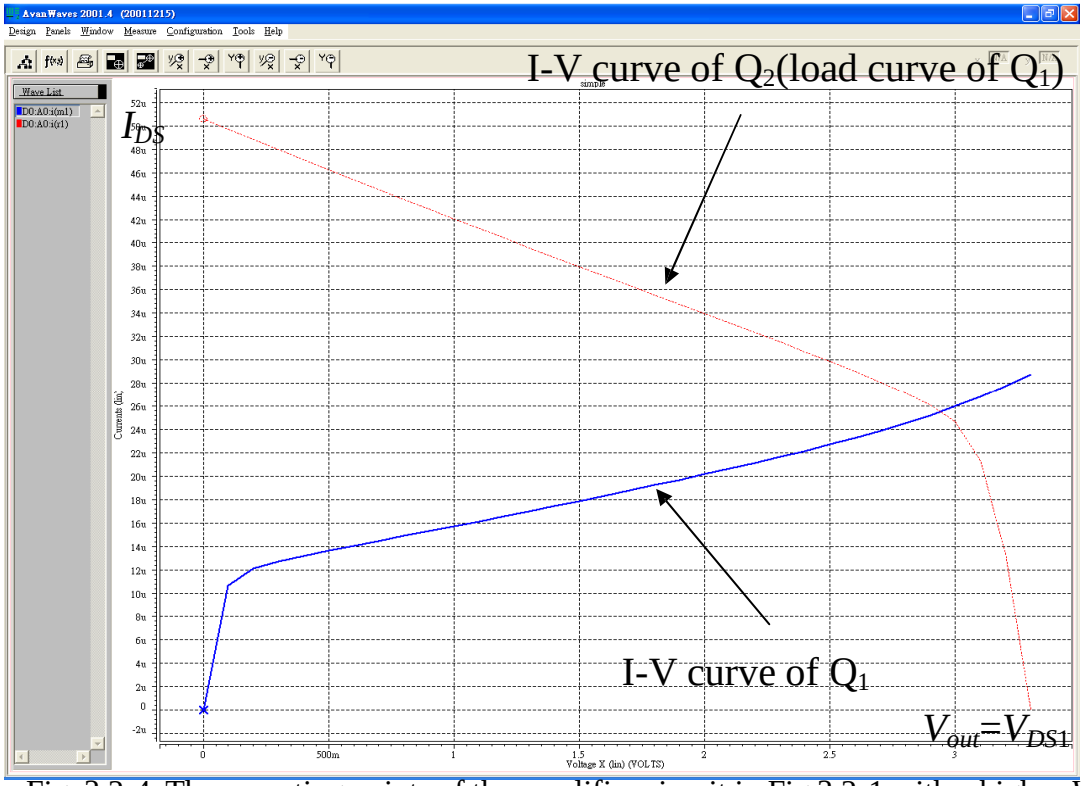


Fig. 3.2-4 The operating points of the amplifier circuit in Fig 3.2-1 with a higher V_{SG2}

From the above experiments, we first conclude that to achieve an appropriate operating point, we must be careful in setting V_{GS1} and V_{SG2}. We also note that the I-V

curves are not so flat as we wished. Therefore, we cannot expect a very high gain with this kind of simple CMOS circuits. As we shall learn in later chapters, the gain can be higher if we use a cascode design.

Experiment 3.2-4 The Gain

We used a signal with magnitude 0.001V and frequency 500kHz. The gain was found to be 30. The program is shown in Table 3.2-4 and the result is shown in Fig. 3.2-5.

Table 3.2-4 Program of Experiment 3.2-4

```
simple
.protect
.lib 'c:\mm0355v.l' TT
.unprotect
.op
.options nomod post
VDD   11      0       3.3v
R1    11      1       0k
VSG    11     2      0.9v

.param W1=5u
M1    3       4       0       0
+nch L=0.35u W='W1' m=1
+AD='0.95u*W1' PD='2*(0.95u+W1)'
+AS='0.95u*W1' PS='2*(0.95u+W1)'
M2    3       2       1       1
+pch L=0.35u W='W1' m=1
+AD='0.95u*W1' PD='2*(0.95u+W1)'
+AS='0.95u*W1' PS='2*(0.95u+W1)'
VGS   4       5       0.65v
Vin   5       0       sin(0 0.001v 500k)

.tran   0.001us         15us

.end
```

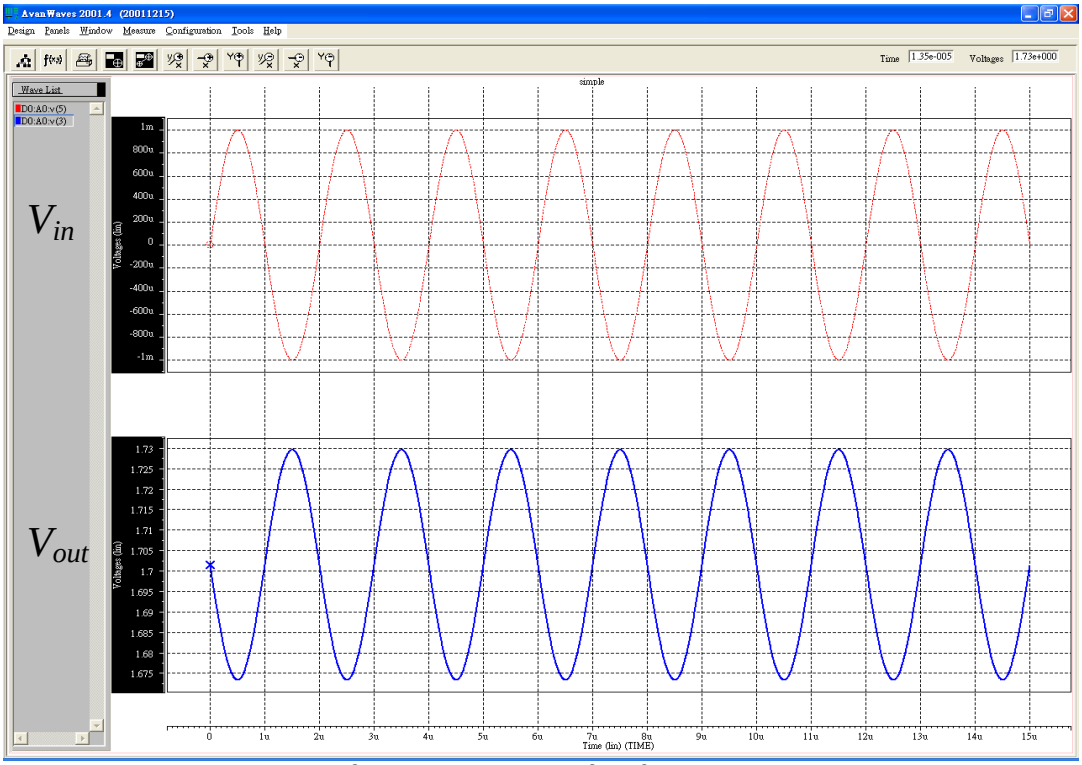


Fig. 3.2-5 The gain of the CMOS amplifier for input signal with 500KHz

Section 3.3 A Desired Current Source

In a CMOS circuit, a V_{SG2} has to be used, as shown in Fig. 3.3-1. In practice, it is not desirable to have many such power supplies all over the integrated circuit. In this section, we shall see how this can be replaced by a desired current source and a current mirror.

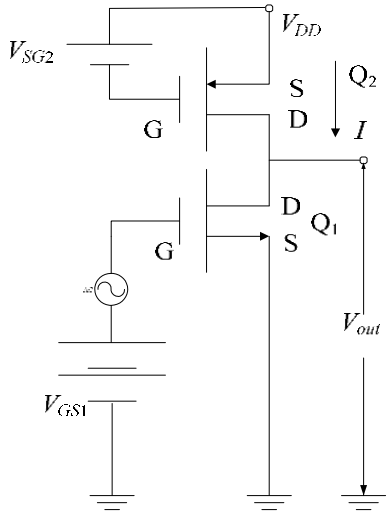
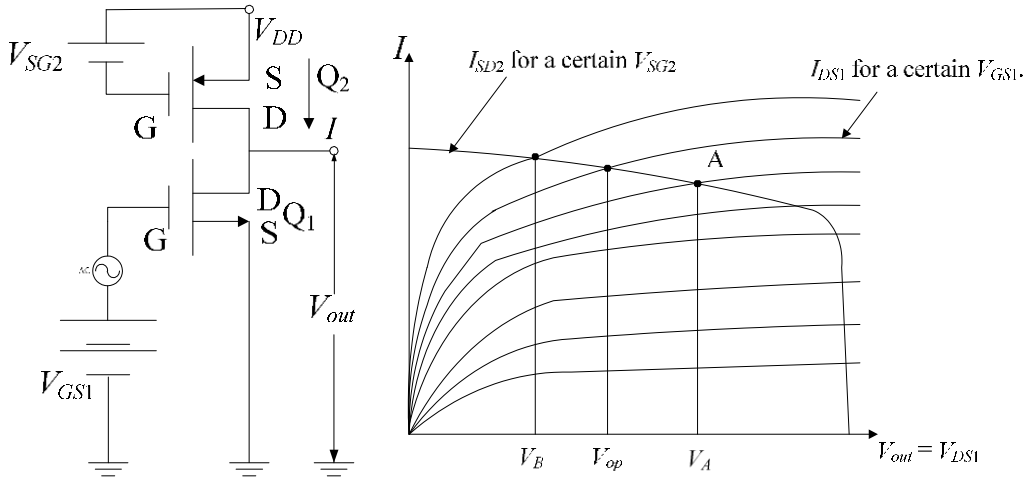


Fig. 3.3-1 A CMOS amplifier with V_{SG2}

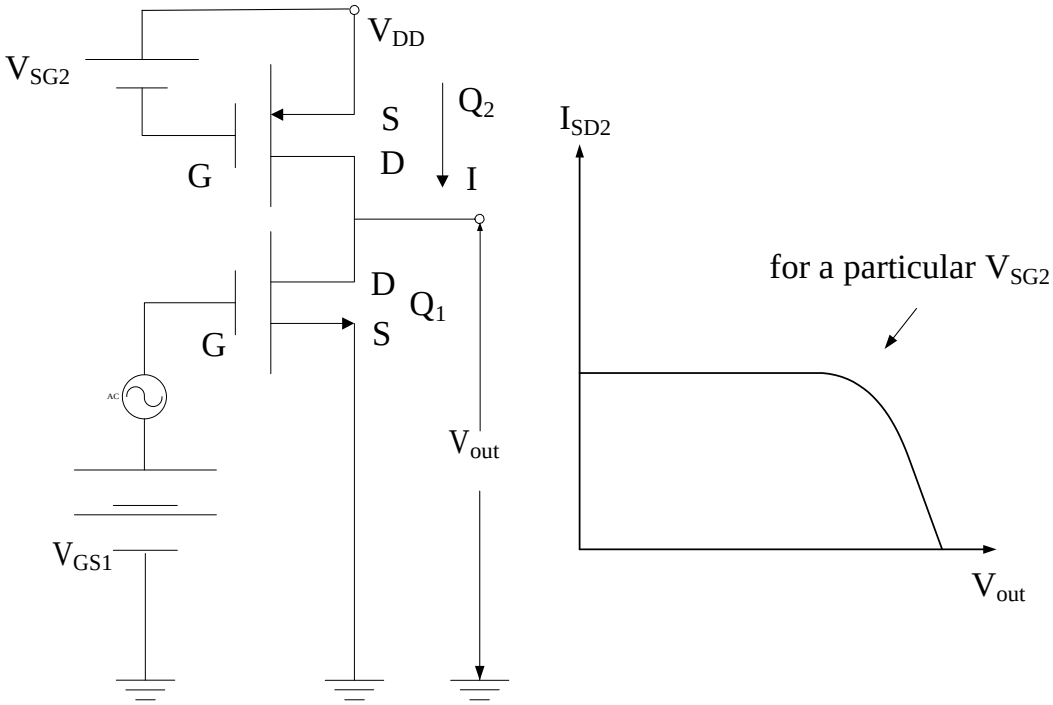
The purpose of V_{SG2} is to produce a desired load curve of Q_1 as shown in Fig. 3.3-2.



(a) A CMOS amplifier with V_{SG2} (b) The I-V currents of Q_1 and its load curve

Fig. 3.3-2 A CMOS amplifier, its I-V curves and load lines

The load curve of Q_1, which corresponds to a particular I-V curve of Q_2, is shown in Fig. 3.3-3. This load curve is determined by V_{SG2}.



(a) A CMOS amplifier with a V_{SG2} (b) I_{SD2} vs V_{out} for the fixed V_{SG2}

Fig. 3.3-3 A CMOS amplifier with a fixed V_{SG2} and its I-V curves

It is natural for us to think that a proper V_{SG2} is the only way to produce the desired load curve for Q_1. Actually, there is another way. Note each load curve almost corresponds to a desired $I_{SD2} = I_{DS1}$, as shown in Fig. 3.3-4. In other words, we may think of a way to produce a desired current in Q_2, which of course is also the current in Q_1.

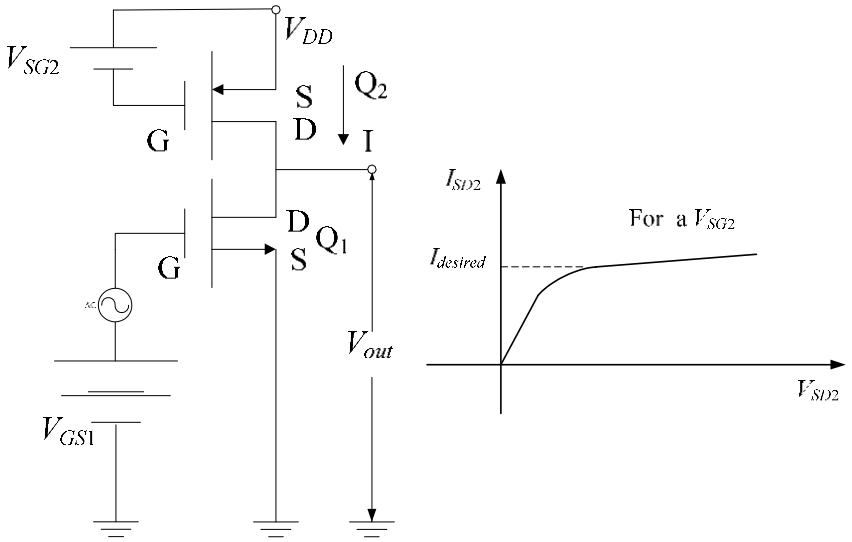


Fig. 3.3-4 An illustration of how a desired current determines the I-V curve

There are two problems here: (1) How can we generate a desired current? (2) How can we force Q_2 to have the desired current?

To answer the first question, let us consider a typical NMOS circuit with a resistive load as shown in Fig. 3.3-5.

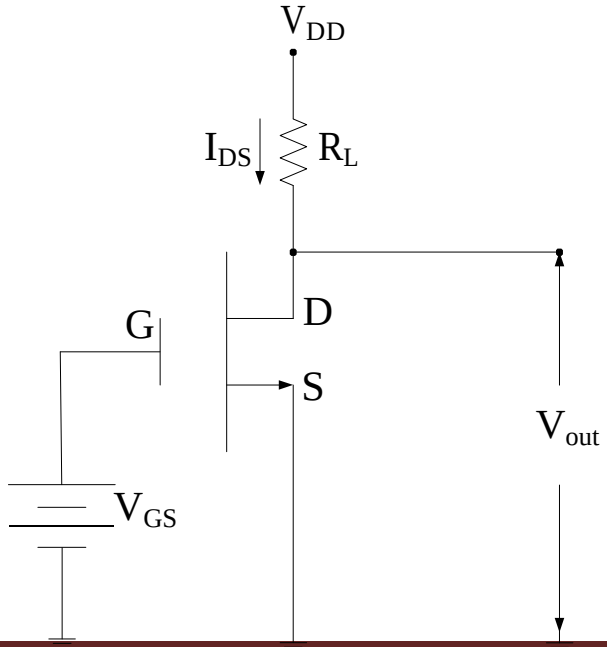


Fig. 3.3-5 An NMOS circuit with a resistive load

In the ohmic region, the relationship between the current I_{DS} and different voltages is expressed as below:

$$I_{DS} = k_n' \frac{W}{L}\left((V_{GS} - V_t)V_{DS} - \frac{1}{2}V_{DS}^2\right) \qquad (3.3\text{-}1)$$

$$I_{DS} = \frac{V_{DD} - V_{DS}}{R_L} \qquad (3.3\text{-}2)$$

Suppose we want to have a desired current I_{DS}. We may think that I_{DS} is a constant. But, from the above equations, we still have three variables, namely V_{GS}, V_{DS} and R_L. Since there are only two equations, we cannot find these three variables for a given desired I_{DS}.

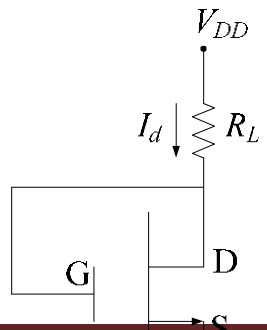
In the boundary between ohmic and saturation regions where $V_{DS} = V_{GS} - V_t$, the two equations governing current and voltages in the transistor are as follows:

$$I_{DS} = \frac{1}{2} k_n' \frac{W}{L}(V_{GS} - V_t)^2 \qquad (3.3\text{-}3)$$

and $$I_{DS} = \frac{V_{DD} - V_{DS}}{R_L} \qquad (3.3\text{-}4)$$

As can be seen, there are still three variables and only two equations.

There is a trick to solve the above problem. We may connect the drain to gate as shown in Fig. 3.3-6.



After this is done, we have

$$V_{GS} = V_{DS} \qquad (3.3\text{-}5)$$

We have successfully eliminated one variable. Besides,

$$V_{GS} - V_t = V_{DS} - V_t \qquad (3.3\text{-}6)$$

From Equation (3.3-6), we have

$$V_{DS} > V_{GS} - V_t \qquad (3.3\text{-}7)$$

Thus, this connection makes sure that the transistor is in saturation region. Since it is in the saturation region, we have

$$I_{DS} = \frac{1}{2} k_n' \left(\frac{W}{L}\right)(V_{GS} - V_t)^2 \qquad (3.3\text{-}8)$$

and $$I_{DS} = \frac{V_{DD} - V_{GS}}{R_L} \qquad (3.3\text{-}9)$$

Although we often say that a transistor is in saturation if its drain is connected to its gate, we must understand it is in a very peculiar situation. Traditionally, a transistor has a family of IV-curves, each of which corresponds to a specified gate bias voltage V_{GS} and besides, the V_{DS} can be any value as illustrated in Fig. 3.3-2. Once the drain is connected to the gate, we note the following:

(1) We have lost V_{DS} because it is always equal to V_{GS}. Therefore, we do not have the traditional IV-curves any more.

(2) For each V_{GS}, since $V_{DS} = V_{GS}$, we have $V_{DS} > V_{GS} - V_t$. This transistor is in saturation. But it is rather close to the boundary between the ohmic region and the saturation region.

(3) Because of the above point, the relationship between current I_{DS} and voltage V_{GS} is the dotted line illustrated in Fig. 3.3-7.

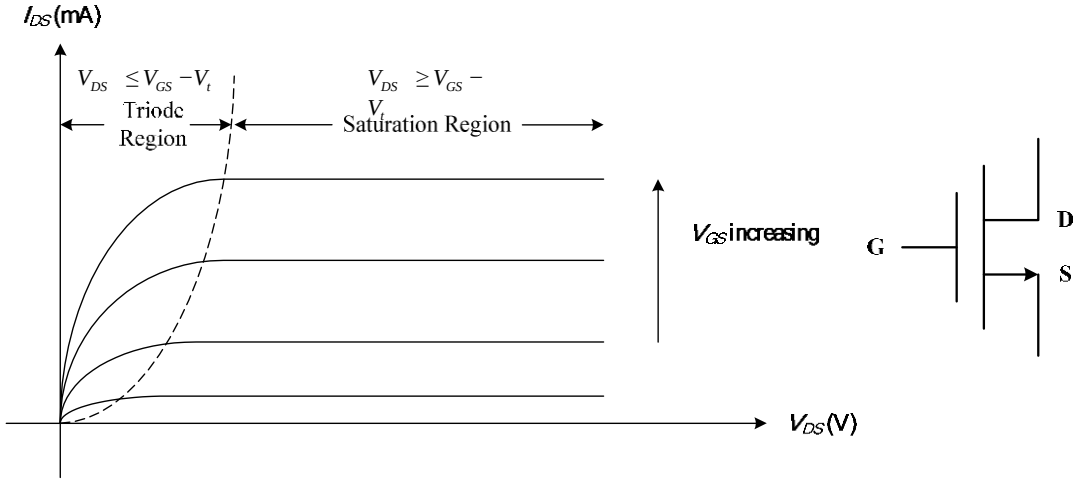
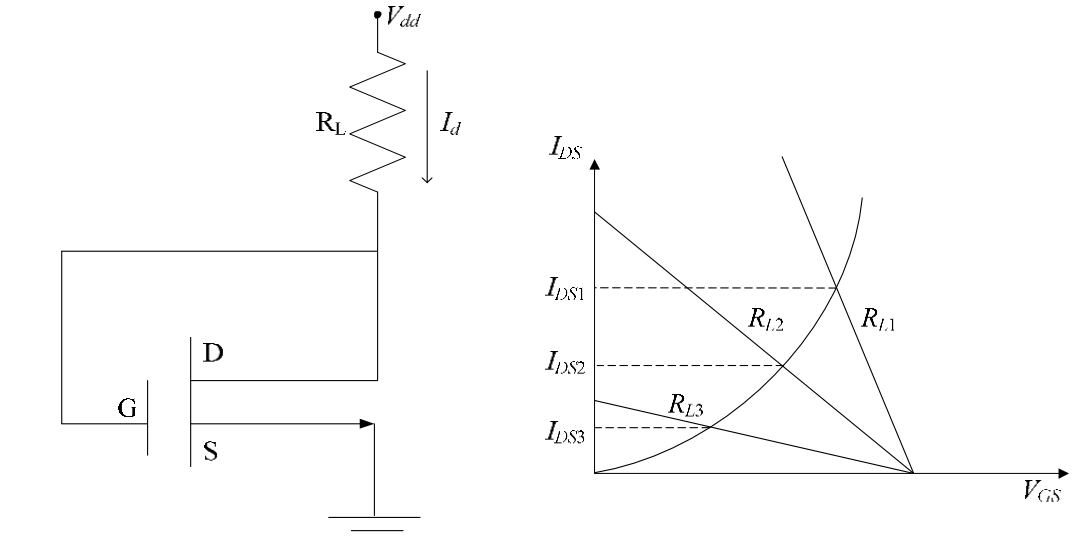


Fig. 3.3-7

(4) We may safely say that the transistor is no longer a transistor. It can be now viewed as a diode with only two terminals. The relationship between current I_{DS} and voltage V_{GS} is hyperbolic expressed in Equation 3.3-8.

(5) For a traditional transistor, V_{GS} is supplied by a bias voltage. Since there is no bias voltage, how do we determine V_{GS}. Note that the desired current is related to V_{GS}. This will be discussed in below.

Given a certain desired I_{DS}, V_{GS} can be determined by using Equations (3.3-8). Thus R_L can be found by using Equation (3.3-9). We can also determine V_{GS} and R_L graphically as shown in Fig. 3.3-8. This means that we can design a desired current source by using the circuit shown in Fig. 3.3-6. By adjusting the value of R_L, we can get the desired current.



(a) A transistor with drain and gate connected (b) The determination of current in a transistor with drain and gate connected

Fig. 3.3-8 The generation of a desired current

Let us examine Fig. 3.3-6 again. We do not have to provide a bias voltage V_{GS} any more. This is a very desirable property which will become clear as we introduce current mirror. But, the reader should note that a V_{GS} does exist and it is produced.

In this section, we have discussed how to generate a desired current. In the next section, we shall show how we can force Q_2 to have this desired current. This is done by the current mirror.

Section 3.4 The Current Mirror

Let us consider the circuit in Fig. 3.4-1.

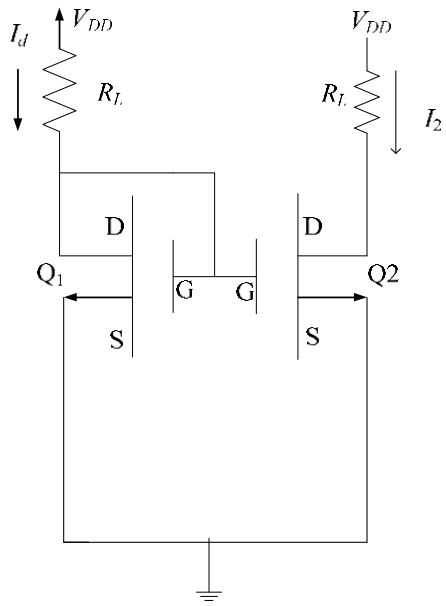


Fig. 3.4-1 A current mirror

Suppose Q_1 and Q_2 have the same V_t. Note that Q_1 is in the saturation region and has a desired current I_d in it. Assume Q_2 is also in the saturation region. Since $V_{GS1} = V_{GS2}$ by using Equation (3.3-3), we have

$$\frac{I_2}{I_d} = \frac{\left(W_2/L_2\right)}{\left(W_1/L_1\right)} \qquad (3.4\text{-}1)$$

If $W_1 = W_2$ and $L_1 = L_2$, from Equation (3.4-1), we have $I_2 = I_d$. Q_1 is called a current mirror for Q_2.

As indicated before, Q_2 must be in the saturation region. So our question is: Under what condition would Q_2 be out of saturation $I_2 = I_d$. Note that Q_2 must be connected to a load. If the load is too high, this will cause it to be out of saturation as illustrated in Fig. 3.4-2.

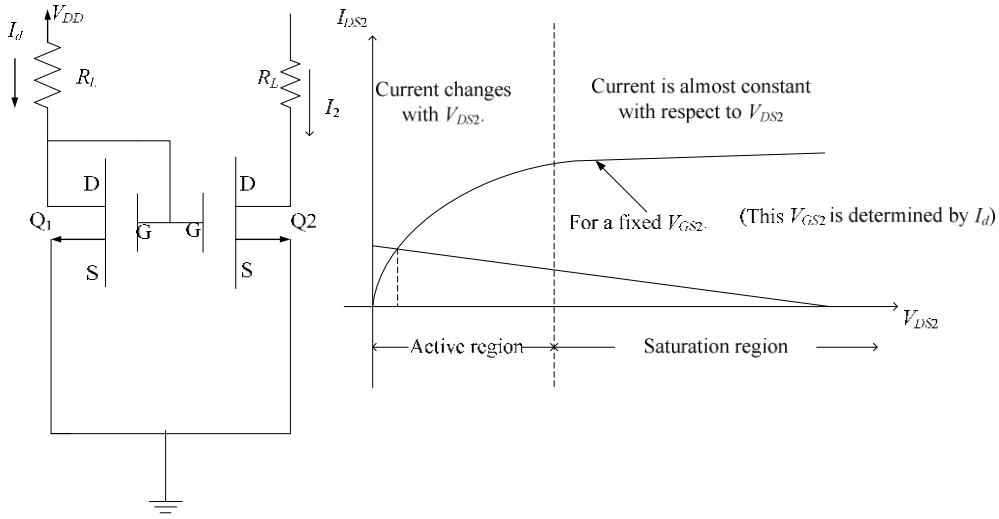


Fig. 3.4-2 The out of saturation of Q_2

The reader may be puzzled about one thing. We know that if an NMOS transistor is in the saturation region, its current is determined by V_{GS}. Is this still true in this case? Our answer is "Yes". That is, for the circuit in Fig. 3.4-1, $I(Q_2)$ is still determined by V_{GS2}. But, we shall now show that V_{GS2} is determined by $I(Q_1)$.

Note that $V_{GS2} = V_{GS1}$. Consider Q_1. The special connection of Q_1 makes $V_{GS1} = V_{DS1}$. But

$$V_{DS1} = V_{DD} - I_{DS1}R_L \qquad (3.4\text{-}2)$$

Thus, from Equation (3.4-2), we conclude that V_{GS2}, which is equal to V_{GS1}, which is in turn equal to V_{DS1}, is determined by $I(Q_1)$.

The advantage of using the current mirror is that no biasing voltage is needed to give a proper V_{GS2}. There is still a V_{GS2}. But this V_{GS2} is equal to V_{GS1} which is in turn determined by $I(Q_1)$. $I(Q_1)$ is determined by selecting a proper R_L, as illustrated in Fig. 3.4-3.

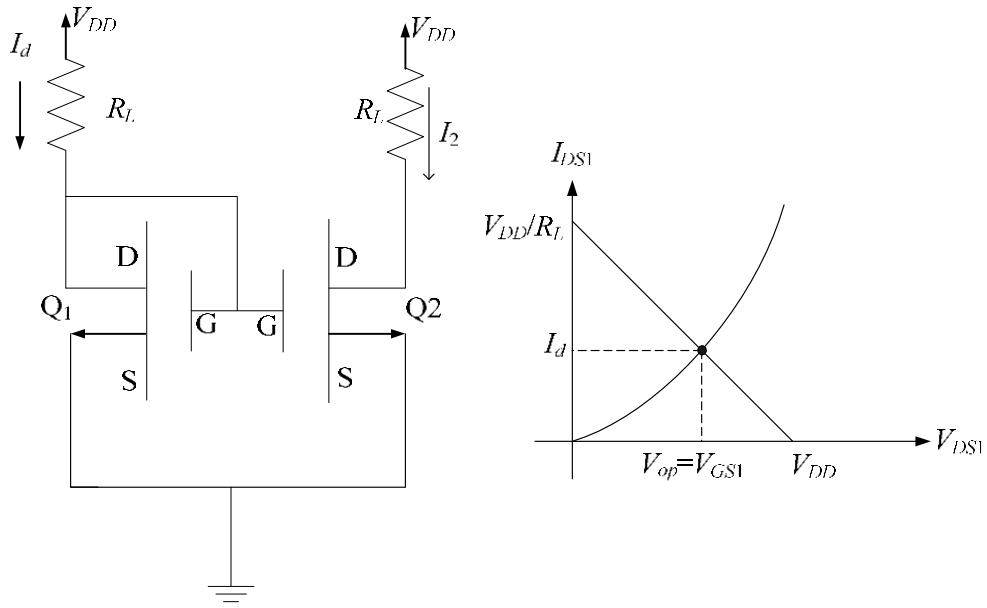


Fig. 3.4-3 The determination of the biasing voltage in a current mirror

A current mirror can be based upon a PMOS transistor as in the CMOS amplifier case. Fig. 3.4-4 shows a CMOS amplifier with a current mirror.

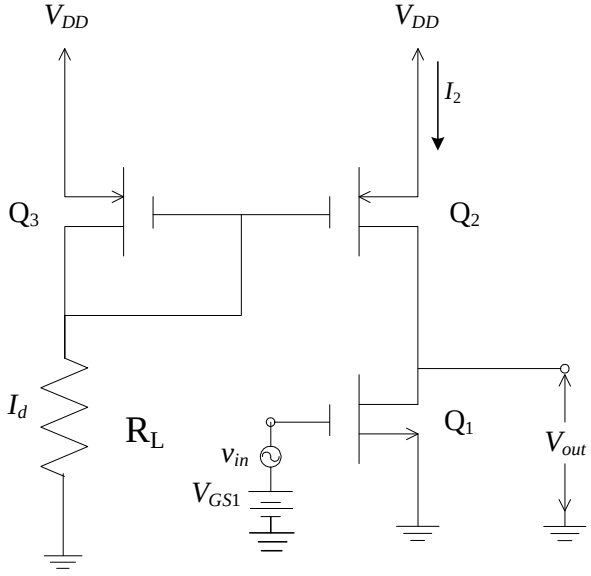
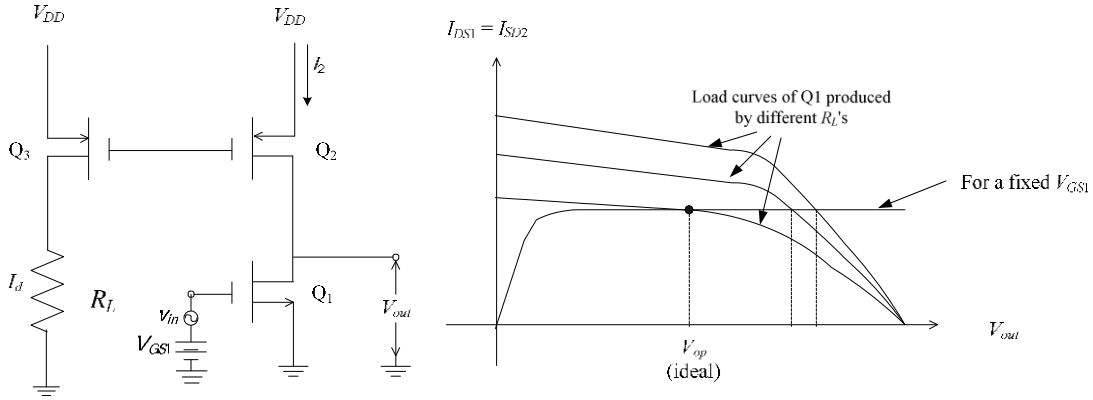


Fig. 3.4-4 A PMOS current mirror

We must remember that the purpose of using a current mirror is to generate a proper I-V curve of Q_2. This I-V curve serves as a load curve for Q_1 as shown in Fig. 3.4-5. From Equation (3.3-8) and (3.3-9), we know that by adjusting the value of R_L, we can obtain different current values in Q_3, which mean different I-V curves in Q_2. In other words, if we want a different load curve of Q_1, we may simply change the value of R_L.



(a) A current mirror (b) Different R_L's producing different I-V curves for Q_1

Fig. 3.4-5 The obtaining of different I-V curves for an NMOS transistor through a current mirror

Section 3.5 Experiments for the CMOS Amplifiers with Current Mirrors

In this set of experiments, we used the circuit shown in Fig. 3.5-1.

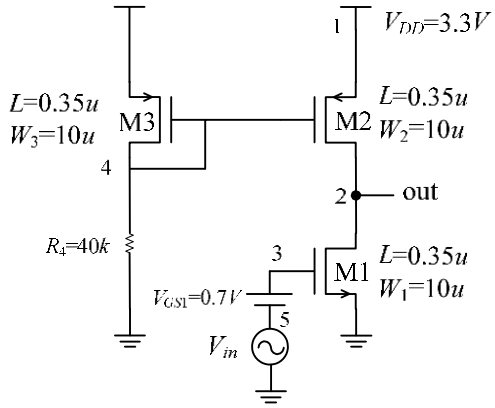


Fig. 3.5-1 The current mirror used in the experiments of Section 3.5

Experiment 3.5-1 The Operating Points of M1 and M3.

In this experiment, we like to find out whether I(M1) is equal to I(M3) or not. We first try to find the characteristics of M1. The program is shown in Table 3.5-1. We then do the same thing to M3. The program is shown in Table 3.5-2. The curves related to M1 are shown in Fig. 3.5-2. The curves related to M3 are shown in Fig. 3.5-3. Note the I-V curve of M3 is not a typical one for a transistor because the gate of M3 is connected to the drain of M3.

Table 3.5-1 Program for
Experiment 3.5-1

```
.lib 'C:\model\tsmc\MIXED035\mm0355v.l' TT
.unprotect
.op
.options nomod post

VDD   1       0       3.3v
R4    4       0       30k
Rdm   1       1_1     0
.param W1=10u  W2=10u  W3=10u  W4=10u
M1    2       3       0       0
+nch L=0.35u  W='W1' m=1   AD='0.95u*W1'
+PD='2*(0.95u+W1)' AS='0.95u*W1' PS='2*(0.95u+W1)'
M2    2       4       1_1     1
+pch L=0.35u
+W='W2' m=1AD='0.95u*W2' PD='2*(0.95u+W2)'
+AS='0.95u*W2' PS='2*(0.95u+W2)'
M3    4       4       1       1
+pch L=0.35u
+W='W3' m=1AD='0.95u*W3' PD='2*(0.95u+W3)'
+AS='0.95u*W3' PS='2*(0.95u+W3)'

V2    2       0       0v
VGS1  3       5       0.7v
Vin   5       0       0v

.DC V2 0 3.3v 0.1v
.PROBE I(M1) I(Rdm)

.end
```

Table 3.5-2 Another program for Experiment 3.5-1

```
Ex3.5-12
.protect
.lib 'c:\mm0355v.l' TT
.unprotect
.op
.options nomod post

VDD   1       0       3.3v
R4    4       0       30k
Rdm   1       1_1     0

.param W1=10u  W2=10u  W3=10u  W4=10u
M1    2       3       0       0
+nch L=0.35u  W='W1' m=1   AD='0.95u*W1'
```

+PD='2*(0.95u+W1)' AS='0.95u*W1' PS='2*(0.95u+W1)'

```
M2      2       4       1       1
+pch L=0.35u
+W='W2' m=1AD='0.95u*W2' PD='2*(0.95u+W2)'
+AS='0.95u*W2' PS='2*(0.95u+W2)'
M3      4       4       1_1     1
+pch L=0.35u
+W='W3' m=1AD='0.95u*W3' PD='2*(0.95u+W3)'
+AS='0.95u*W3' PS='2*(0.95u+W3)'

V3      4       0       0v
VGS1    3       5       0.7v
Vin     5       0       0v

.DC V3 0 3.3v 0.1v
.PROBE I(R4) I(Rdm)
.end
```

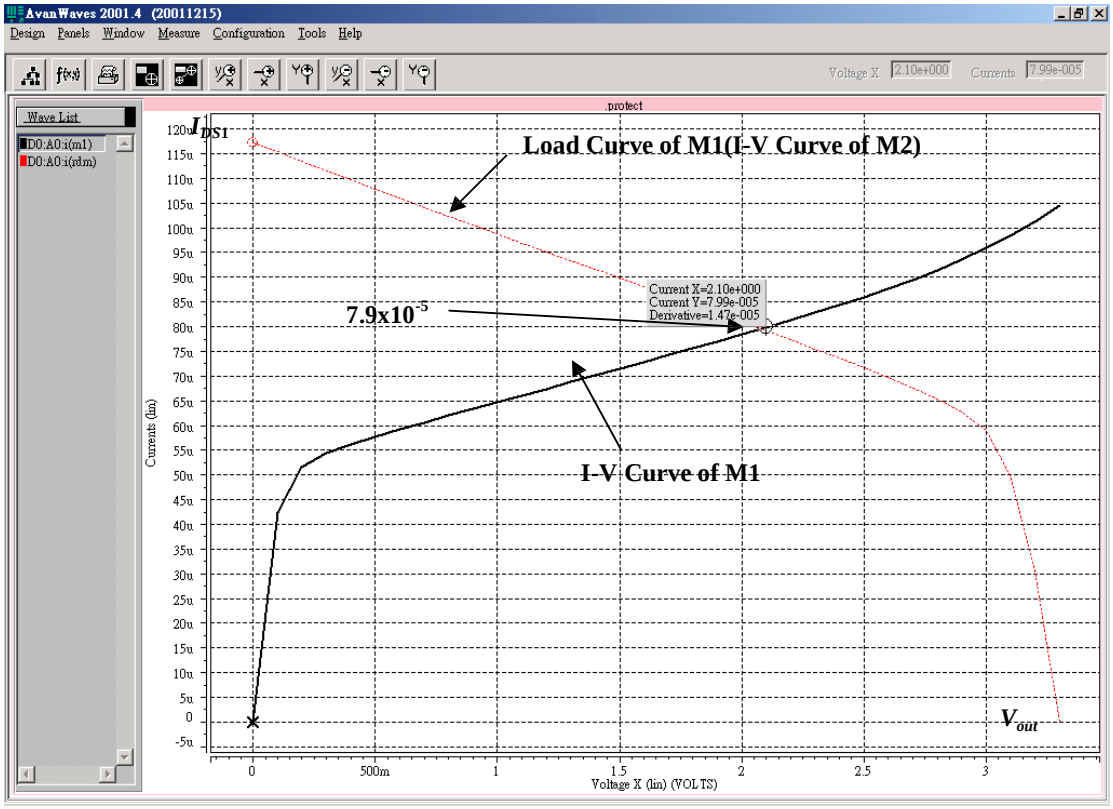


Fig. 3.5-2 I-V curve and load curve for M1

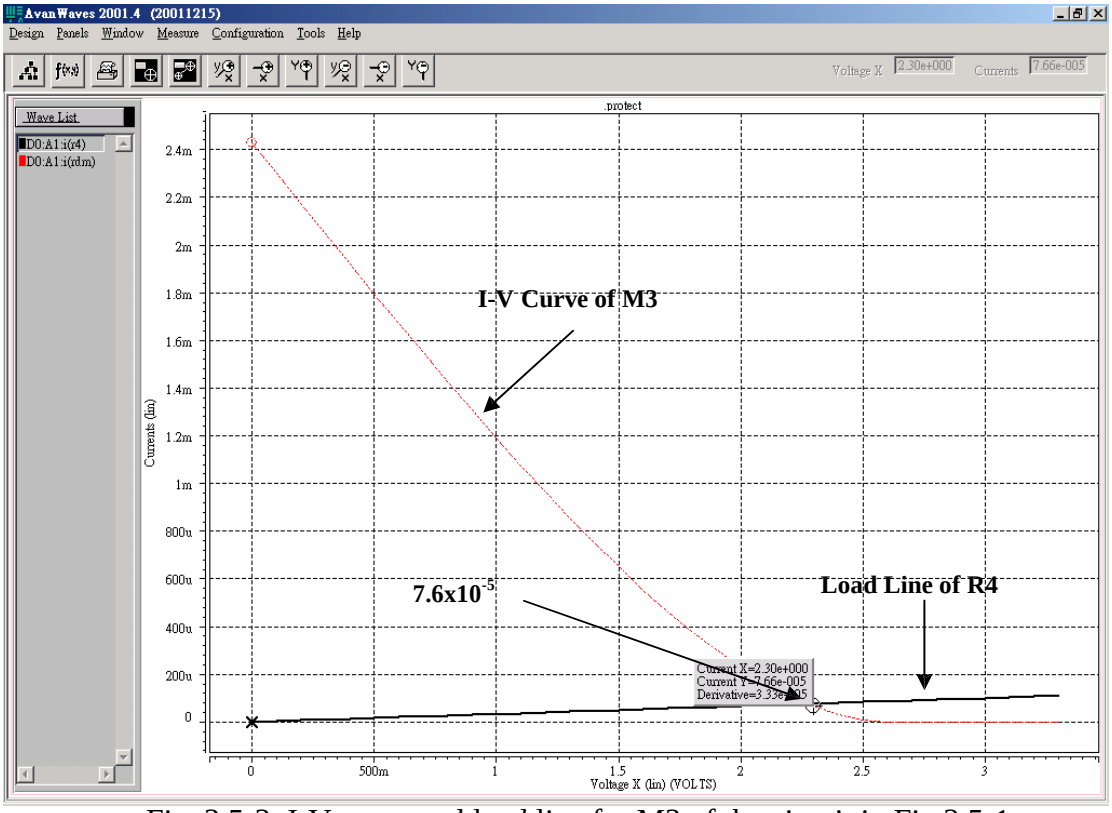


Fig. 3.5-3 I-V curve and load line for M3 of the circuit in Fig 3.5-1

From this experiment, we conclude that I(M3)=I(M1) as expected.

Experiment 3.5-2 The Operating Point of M2

The I-V curve of M2 is the load curve of M1. The I-V curve of M2 is determined by the current mirror mechanism. We were told that the current mirror works only when M2 is in the saturation region. In this experiment, we first show the characteristics of M1. The program is shown in Table 3.5-3. The I-V curve of M2 and its load curve (M1 is the load of M2) are shown in Fig. 3.5-4.

Table 3.5-3 Program for Experiment 3.5-2

```
Ex3.5-2
.protect
.lib 'c:\mm0355v.l' TT
.unprotect
.op
.options nomod post

VDD  1     0     3.3v
R4   4     0     30k
```

.param W1=10u W2=10u W3=10u W4=10u

```
M1      2       3       0       0
+nch L=0.35u W='W1' m=1  AD='0.95u*W1'
+PD='2*(0.95u+W1)' AS='0.95u*W1' PS='2*(0.95u+W1)'
M2      2       4       1_1     1
+pch L=0.35u
+W='W2' m=1AD='0.95u*W2' PD='2*(0.95u+W2)'
+AS='0.95u*W2' PS='2*(0.95u+W2)'
M3      4       4       1       1
+pch L=0.35u
+W='W3' m=1AD='0.95u*W3' PD='2*(0.95u+W3)'
+AS='0.95u*W3' PS='2*(0.95u+W3)'

V2      2       0       0v
VGS1    3       5       0.7v
Vin     5       0       0v

.DC V2 0 3.3v 0.1v
.PROBE I(M1) I(Rdm)
Rdm     1       1_1     0

.end
```

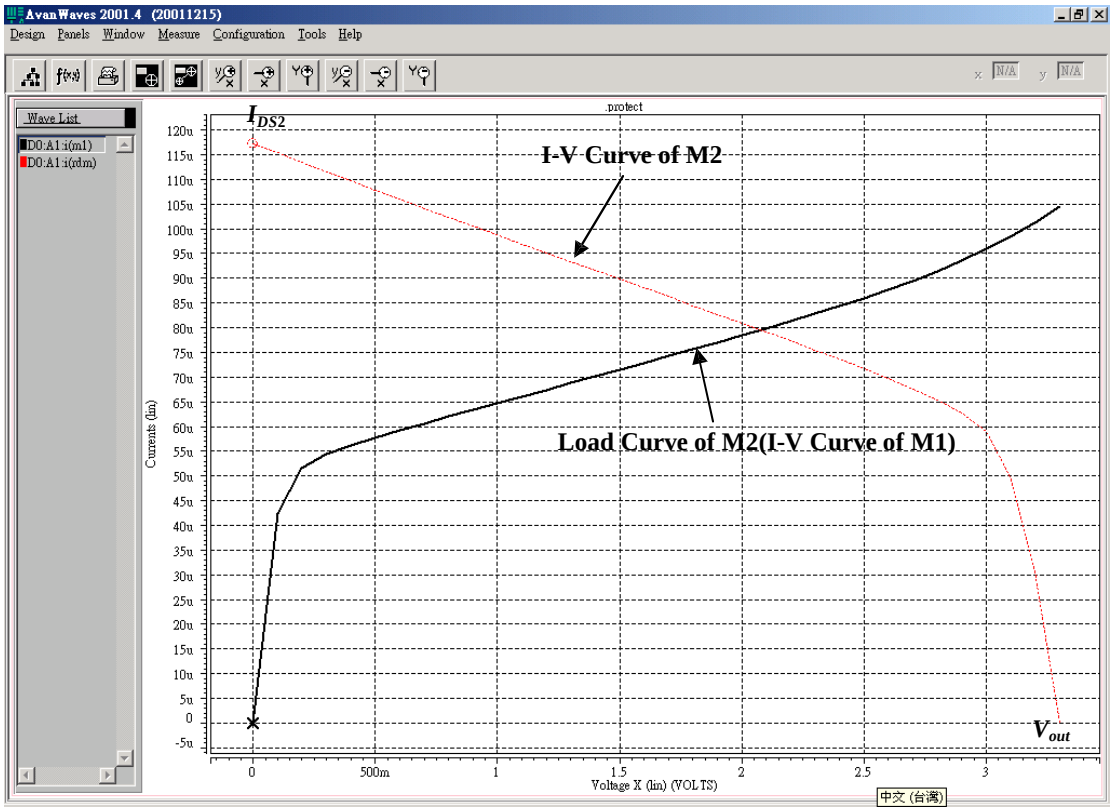


Fig. 3.5-4 Operating points of M2 of the circuit in Fig 3.5-1

As shown in Fig. 3.5-4, M2 is in the saturation region.

To drive M2 out of the saturation region, we lowered V_{GS1} from 0.7V to 0.6V. The program is shown in Table 3.5-4 and the curves are shown in Fig. 3.5-5.

Table 3.5-4 The program to drive M2 out of saturation

```
Ex3.5-2b
.protect
.lib 'c:\mm0355v.l' TT
.unprotect
.op
.options nomod post

VDD   1       0       3.3v
R4    4       0       30k

.param W1=10u  W2=10u  W3=10u  W4=10u
M1    2       3       0       0
+nch L=0.35u  W='W1' m=1   AD='0.95u*W1'
+PD='2*(0.95u+W1)' AS='0.95u*W1' PS='2*(0.95u+W1)'
M2    2       4       1_1     1
+pch L=0.35u
+W='W2' m=1AD='0.95u*W2' PD='2*(0.95u+W2)'
+AS='0.95u*W2' PS='2*(0.95u+W2)'
M3    4       4       1       1
+pch L=0.35u
+W='W3' m=1AD='0.95u*W3' PD='2*(0.95u+W3)'
+AS='0.95u*W3' PS='2*(0.95u+W3)'

V2    2       0       0v
VGS1  3       5       0.6v
Vin   5       0       0v

.DC V2 0 3.3v 0.1v
.PROBE I(M1) I(Rdm)
Rdm   1       1_1     0

.end
```

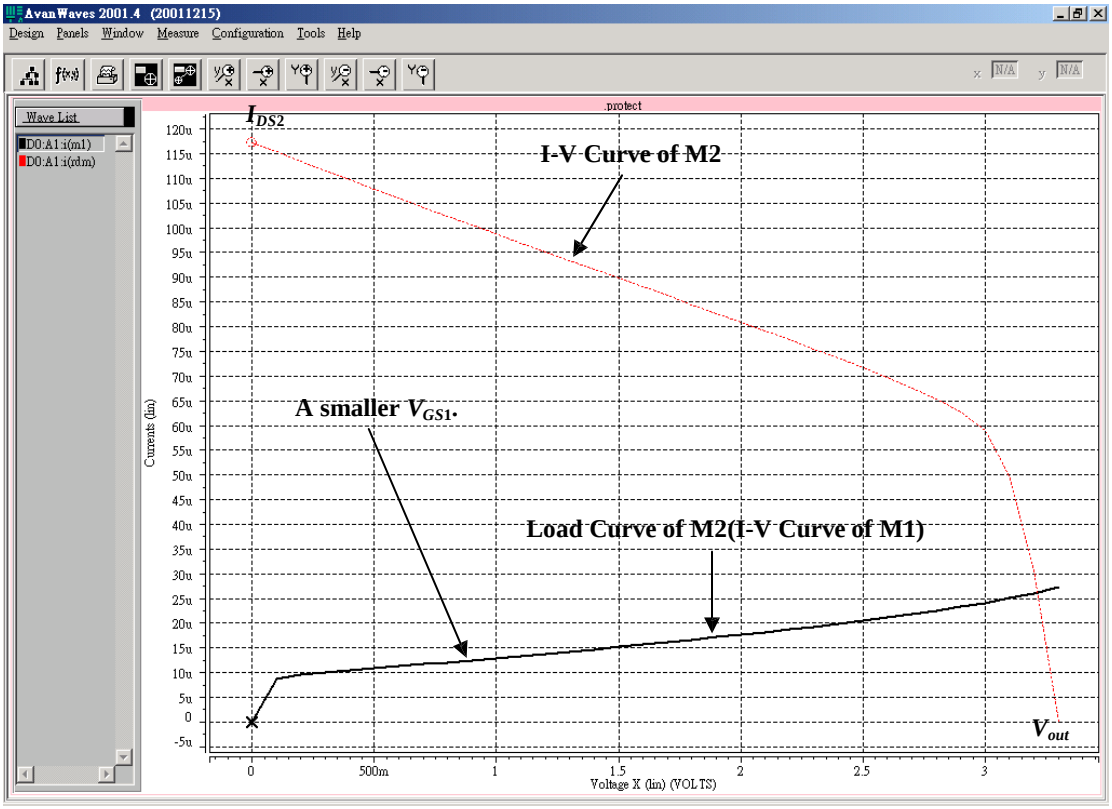


Fig. 3.5-5 The out of saturation of M2

From Fig. 3.5-5, we can see that M2 is now out of saturation. We then printed the essential data by using the SPICE simulation program in Table 3.5-5. We can see that I(M2) is quite different from I(M3) now. This is due to the fact that M2 is out of saturation.

Table 3.5-5 Experimental data for Experiment 3.5-2

subckt			
element	0:m1	0:m2	0:m3
model	0:nch.3	0:pch.3	0:pch.3
region	Saturati	Linear	Saturati
id	25.4028u	-26.2672u	-75.8881u
ibs	-3.880e-17	6.373e-18	1.832e-17
ibd	-864.4522n	1.0916f	1.1504f
vgs	600.0000m	-1.0234	-1.0234
vds	3.2166	-83.3759m	-1.0234
vbs	0.	0.	0.
vth	545.8793m	-719.8174m	-688.8560m
vdsat	85.3814m	-293.2779m	-318.3382m
beta	6.7003m	1.3100m	1.3166m
gam eff	591.1171m	485.8319m	485.8388m
gm	441.4749u	97.6517u	411.0201u
gds	8.7037u	268.5247u	18.5395u

gmb 111.6472u 22.6467u 87.7975u

cdtot	11.3338f	28.6456f	14.4251f
cgtot	10.2012f	16.2693f	12.7341f
cstot	21.1660f	31.1152f	30.5144f
cbtot	27.1028f	38.9009f	33.8541f
cgs	5.6263f	8.8368f	9.9096f
cgd	2.0774f	7.2706f	1.8392f

Experiment 3.5-3 The DC Input-Output Relationship of M1.

In this experiment, weplotted V_{DS1} versus V_{GS1}. The program is in Table 3.5-6 and the DC input-output relationship is shown in Fig. 3.5-6.

Table 3.5-6 Program of Experiment 3.5-3

```
.protect
.lib 'c:\mm0355v.l' TT
.unprotect
.op
.options nomod post

VDD   1        0        3.3v
R4    4        0        30k

.param W1=10u  W2=10u  W3=10u  W4=10u
M1    2        3        0        0
+nch L=0.35u W='W1' m=1  AD='0.95u*W1'
+PD='2*(0.95u+W1)' AS='0.95u*W1' PS='2*(0.95u+W1)'
M2    2        4        1        1
+pch L=0.35u
+W='W2' m=1AD='0.95u*W2' PD='2*(0.95u+W2)'
+AS='0.95u*W2' PS='2*(0.95u+W2)'
M3    4        4        1        1
+pch L=0.35u
+W='W3' m=1AD='0.95u*W3' PD='2*(0.95u+W3)'
+AS='0.95u*W3' PS='2*(0.95u+W3)'

VGS1  3        0        0v

.DC VGS1 0 3.3v 0.1v
.PROBE I(M1)

.end
```

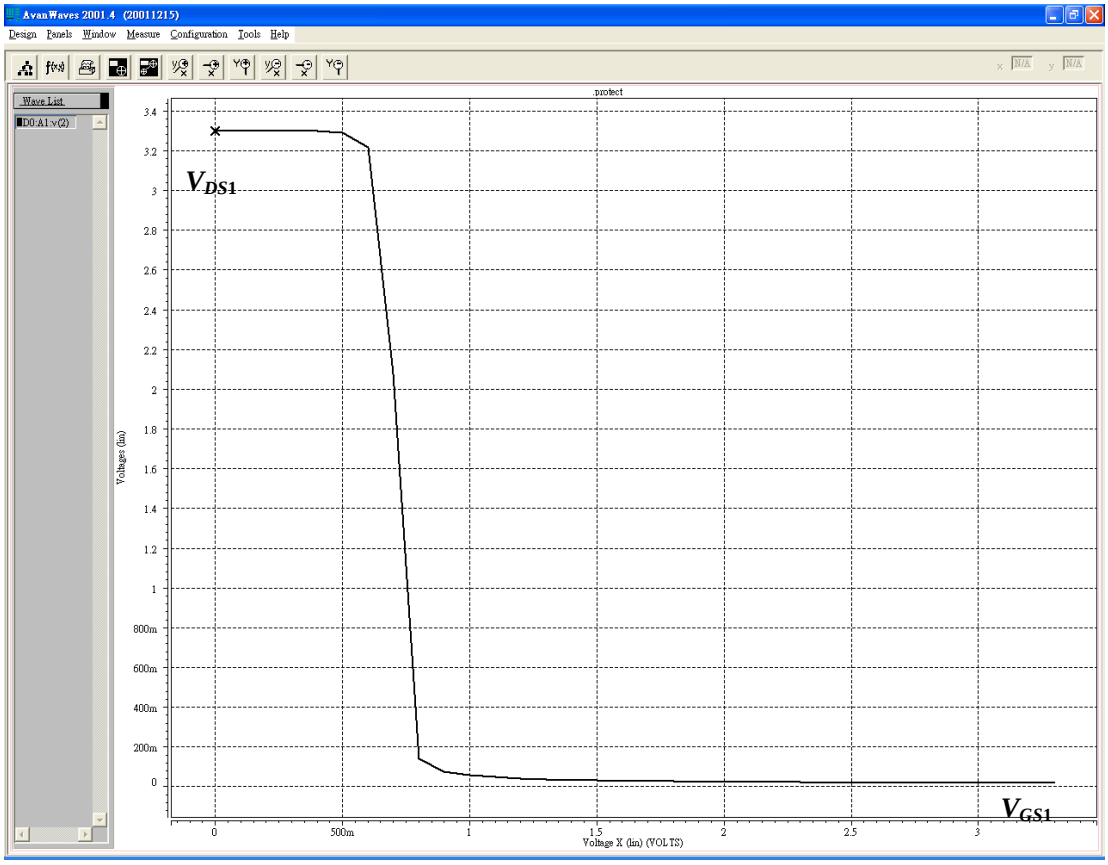
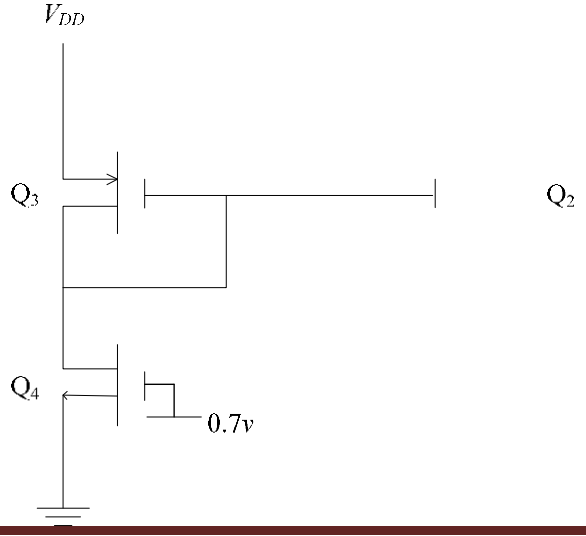


Fig. 3.5-6 V_{DS1} vs V_{GS1}

Section 3.6 The Current Mirror with an Active Load

In the above sections, the current mirror has a resistive load. As we indicated before, a resistive load is not practical in VLSI design. Therefore, it can be replaced by an active load, namely a transistor. Fig. 3.6-1 shows a typical CMOS amplifier whose current mirror has an active load.



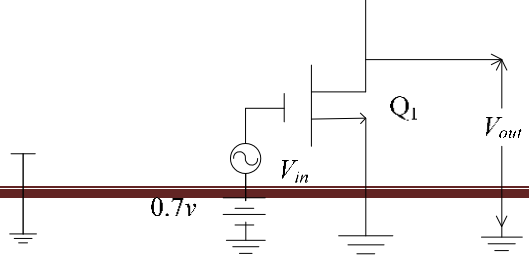

Q1
V_{out}
V_{in}
0.7v

Fig. 3.6-1 A current mirror with an active load

In the above circuit, Q_3 is a current mirror while Q_4 is its load. Note that the main purpose of having a current mirror is to produce a desired basing current in Q_2 which is equal to the current in Q_3. To generate such a desired current, we use the I-V curve of Q_3 and its load curve, which is the I-V curve of Q_4. These curves are shown in Fig. 3.6-2.

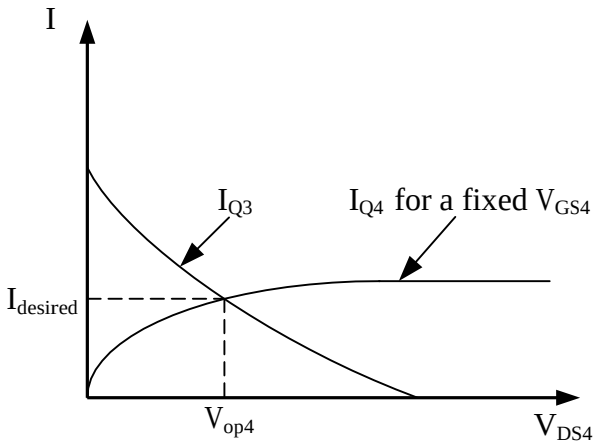


Fig. 3.6-2 The determination of operating point for M4 in Fig. 3.6-1

Note that we have a desired current in our mind. So we just have to adjust V_{GS4} such that its corresponding I-V curve intersects the I-V curve of Q_3 at the proper place which gives us the desired current in Q_4, which is also the current in Q_3.

We indicated before that we like to use current mirrors because we do not like to have two biases as required in a CMOS circuit shown in Fig. 3.1-5. One may wonder at this point that we need two power supplies (constant voltage sources) for this current mirror circuit in the circuit shown in Fig. 3.6-1. Note that in this circuit, although there are two biases, they can be designed to be the same. Thus, actually, we only need one bias. If no current mirror is used in a CMOS circuit, we must need two different biases.

Besides, it will be shown in the next chapter that the current mirror actually has an entirely different function. That is, it provides a feedback in the differential amplifier which gives us a high gain.

Section 3.7 Experiments with the Current Mirror with an Active Load

In the experiments, we used the amplifier circuit shown in Fig. 3.7-1.

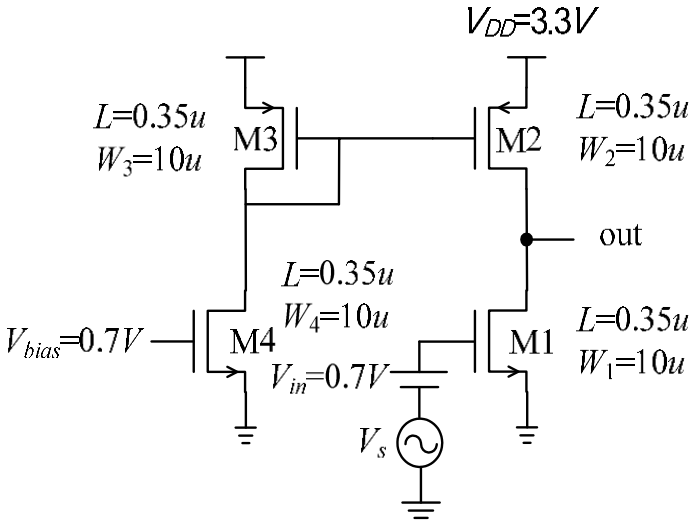


Fig. 3.7-1 The current mirror circuit for experiments in Section 3.7

Experiment 3.7-1 The Operating Point of M4.

The program for this experiment is shown in Table 3.7-1. The curves are shown in Fig. 3.7-2. We like to point out again that the load curve of M4 is the I-V curve of M3. This I-V curve of M3is hyperbola because the gate of M3 is connected to the drain of M3. The result shows that the current is 100u, a quite small value.

Table 3.7-1 Program for Experiment 3.7-1

```
.protect
.lib 'c:\mm0355v.l' TT
.unprotect
.op
.options nomod post

VDD     1       0       3.3v

.param   W1=10u  W2=10u  W3=10u  W4=10u
M1       2       3       0       0
+nch L=0.35u    W='W1' m=1   AD='0.95u*W1'
+PD='2*(0.95u+W1)' AS='0.95u*W1' PS='2*(0.95u+W1)'
M2       2       4       1_1     1
+pch L=0.35u
+W='W2' m=1   AD='0.95u*W2' PD='2*(0.95u+W2)'
+AS='0.95u*W2' PS='2*(0.95u+W2)'
M3       4       4       3_1     1
+pch L=0.35u
```

```
+W='W3' m=1   AD='0.95u*W3' PD='2*(0.95u+W3)'
+AS='0.95u*W3' PS='2*(0.95u+W3)'
M4        4        5        0        0
+nch L=0.35u
+W='W4' m=1   AD='0.95u*W4' PD='2*(0.95u+W4)'
+AS='0.95u*W4' PS='2*(0.95u+W4)'

V4     4      0      0v
VGS1   3      6      0.7v
VGS4   5      0      0.7v
Vin    6      0      0v
.DC V4 0 3.3v 0.1v
.PROBE I(M4) I(Rm3)
Rdm      1      1_1    0
Rm3      1      3_1    0

.end
```

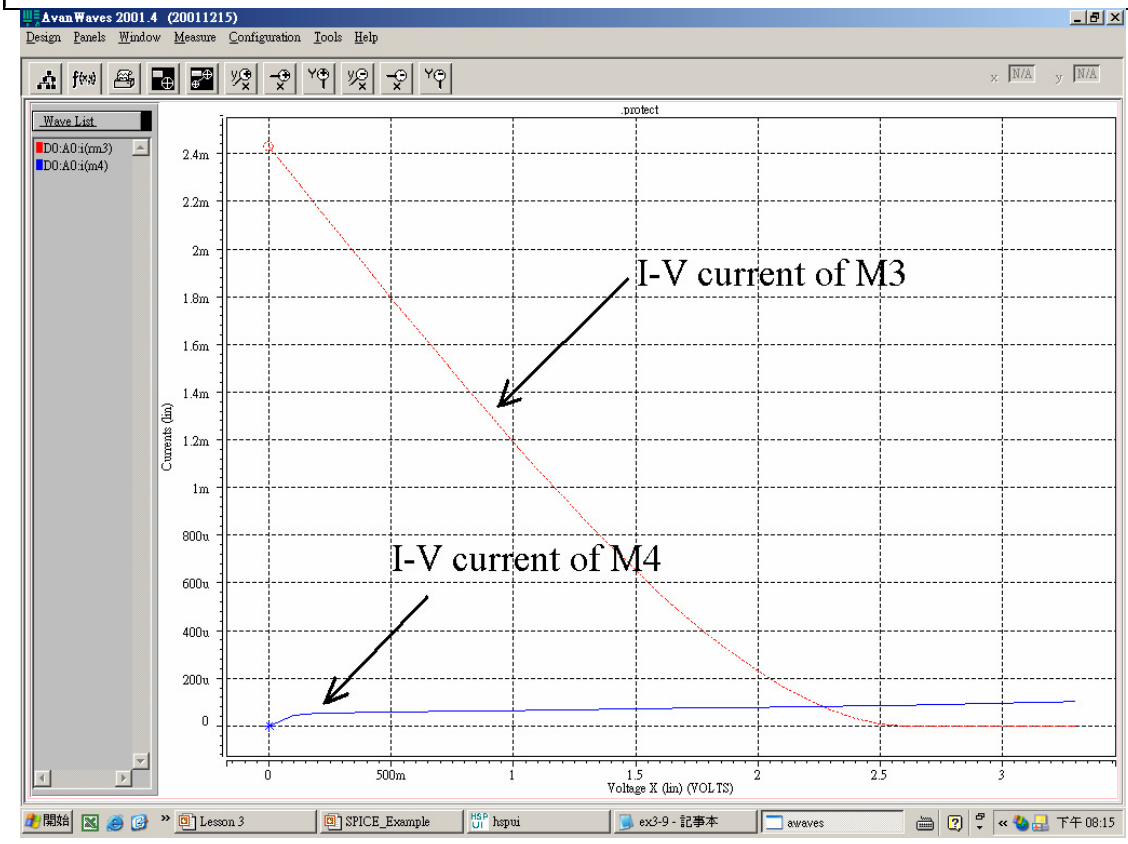


Fig. 3.7-2 Operating point of M4

Section 4.5 The Small Signal Analysis of the Differential Amplifier with Active Loads

Let us redraw the differential amplifier circuit in Fig. 4.5-1.

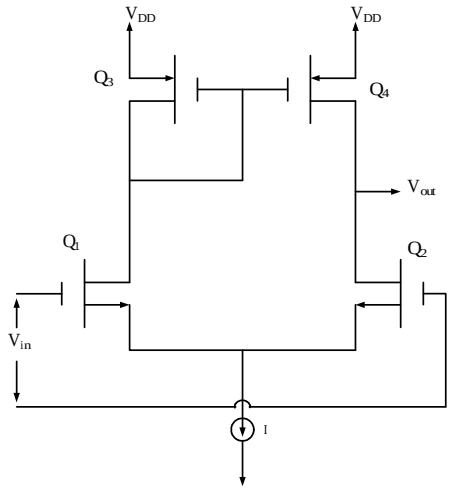


Fig. 4.5-1 A differential amplifier with active loads for AC analysis

To find its small signal equivalent circuit, we first note that Q_3 is a specially connected PMOS transistor. Its small signal equivalent circuit, as discussed in Section 4.3, is now shown in Fig. 4.5-2.

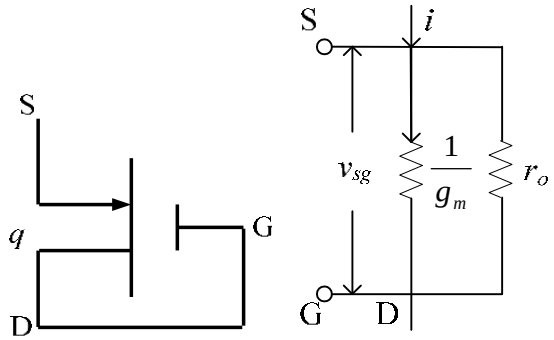


Fig. 4.5-2 A small signal equivalent circuit for a PMOS transistor with gate and drain connected together

The small signal equivalent circuit of the differential amplifier with active loads is shown in Fig. 4.5-3.

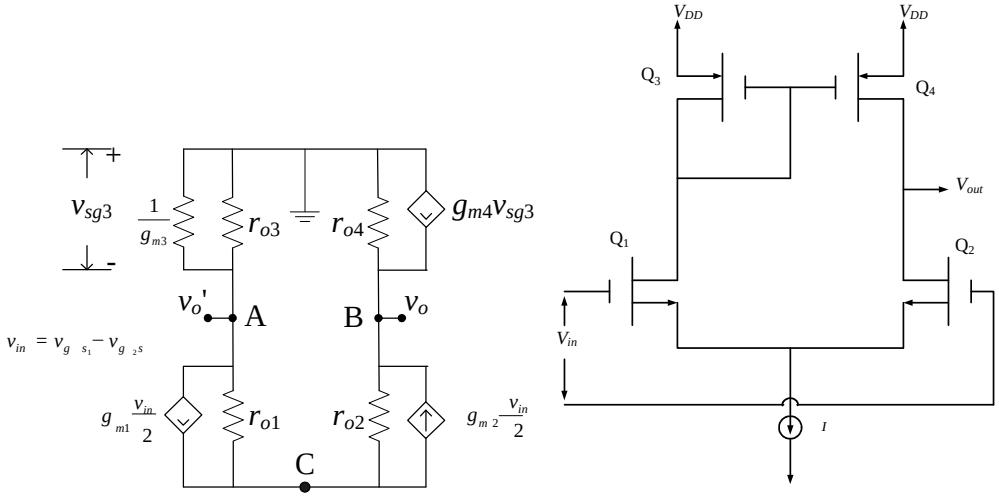


Fig. 4.5-3 The small signal equivalent circuit for the circuit in Fig. 4.5-1

Consider Node A. We have

$$\frac{v_0'-v_s}{r_{01}} + \frac{g_{m1}}{2}v_{in} + \frac{v_0'}{\frac{1}{g_{m3}} // r_{01}} = 0 \qquad (4.5\text{-}1)$$

Since r_{01} is much larger than $1/g_{m3}$, we have

$$\frac{v_0'-v_s}{r_{01}} + \frac{g_{m1}}{2}v_{in} + g_{m3}v_0' = 0 \qquad (4.5\text{-}2)$$

Consider Node B.

$$\frac{v_0}{r_{04}} + \frac{v_0 - v_s}{r_{02}} = g_{m4}v_{sg3} + g_{m2}\frac{v_m}{2}$$

Since $v_{sg3} = -v_0'$, we have:

$$\frac{v_0}{r_{04}} + \frac{v_0 - v_s}{r_{02}} = -g_{m4}v_0' + g_{m2}\frac{v_m}{2} \qquad (4.5\text{-}3)$$

Consider Node C.

$$\frac{v_0 - v_0'}{g} + \frac{v_s - v_0}{} = \frac{v_{in}}{} - g\ \frac{v_{in}}{} \qquad (4.5\text{-}4)$$

r_{01} r_{02} ^{m1} 2 ^{m2} 2

To simplify the discussion, we assume that $g_{m1} = g_{m2} = g_{m3} = g_{m4} = g_m$ and $r_{01} = r_{02} = r_{03} = r_{04} = r_0$. Thus, we have

$$\frac{v_0' - v_s}{r_0} + \frac{g_m}{2} v_{in} + g_m v_0' = 0 \qquad (4.5\text{-}5)$$

$$\frac{v_0}{r_0} + \frac{v_0 - v_s}{r_0} + g_m v_0' - g_m \frac{v_m}{2} = 0 \qquad (4.5\text{-}6)$$

$$\frac{v_0 - v_0'}{r_0} + \frac{v_s - v_0}{r_0} = 0 \qquad (4.5\text{-}7)$$

(4.5-5)+(4.5-6)+(4.5-7):

$$2g_m v_0' + \frac{v_0}{r_0} = 0$$

$$v_0' = \frac{v_0}{2g_m r_0} \qquad (4.5\text{-}8)$$

$$v_s = \frac{v_0 + v_0'}{2}$$

$$v_s = \frac{1}{2}(v_0 - \frac{v_0}{2g_m r_0}) \qquad (4.5\text{-}9)$$

Substituting (4.5-8) and (4.5-9) into (4.5-5), we have:

$$\frac{1}{r_0}(\frac{-v_0}{2g_m r_0}) - \frac{v_0}{2r_0} + \frac{v_0}{4g_m r_0^2} + g_m(\frac{-v_0}{2g_m r_0}) + g_m \frac{v_{in}}{2} = 0 \qquad (4.5\text{-}10)$$

$$v_0(\frac{1}{4g_m r_0^2} + \frac{1}{2r_0} + \frac{1}{2r_0}) = \frac{g_m}{2} v_{in} \qquad (4.5\text{-}11)$$

$$v_0(\frac{1 + 4g_m r_0}{4g_m r_0^2}) = \frac{g_m}{2} v_{in} \qquad (4.5\text{-}12)$$

Since $4g_m r_0 >> 1$, we have

$$\text{Gain} = \frac{v_0}{v_{in}} = \frac{1}{2} g_m r_0 \qquad (4.5\text{-}13)$$

Thus the gain is quite large.

Let us find v_0'. Using Equations (4.5-13) and (4.5-8), we have:

$$v_0' = -\frac{v_0}{2 g_m r_o} = \frac{1}{4} \frac{g_m r_0}{g_m r_0} v_{in} = \frac{1}{4} v_{in} \qquad (4.5\text{-}14)$$

Note that v_{in} is very small. The above equation confirms our statement made before that the small signal voltage at the drain of M_3 can be ignored.

Cascode Current Mirror

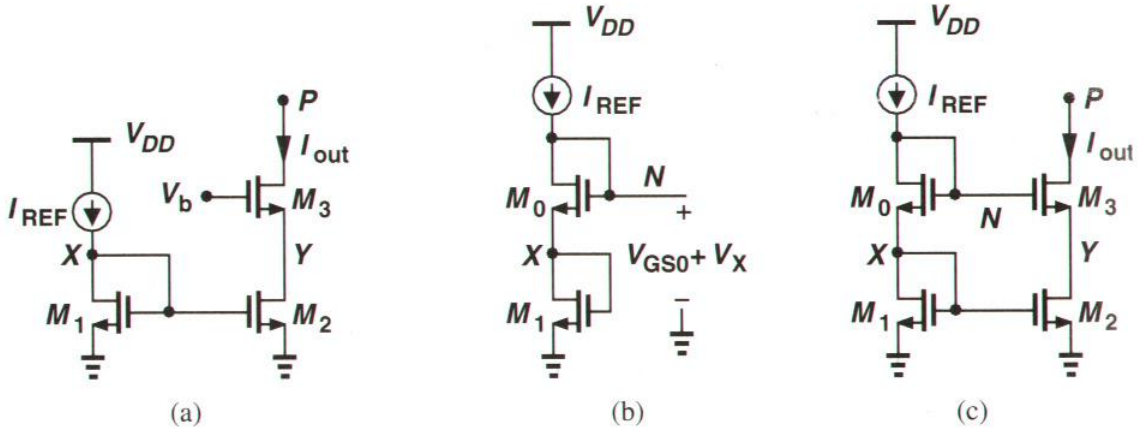


Figure 5.9 (a) Cascode current source, (b) modification of mirror circuit to generate the cascode bias voltage, (c) cascode current mirror.

- In order to suppress the effect of channel-length modulation, a cascode current source can be used.
- If V_b is chosen such that $V_y = V_x$, then I_{out} closely tracks I_{REF}. This is because the cascode device "**shields**" the bottom transistor from variations in V_P. Remember, as long as the drain current is constant, the drain voltage will not change.
- While L_1 must be equal to L_2, the length of M_3 need not be equal to L_1 and L_2.
- To ensure $V_Y = V_X$, we must ensure that

  $\boldsymbol{V_b - V_{GS3} = V_Y}$ or $\boldsymbol{V_b - V_{GS3} = V_X}$ or $\boldsymbol{V_b = V_{GS3} + V_X}$

  This can be done by adding another diode-connected device M_0 that will have

  $\boldsymbol{V_b = V_{GS0} + V_X}$

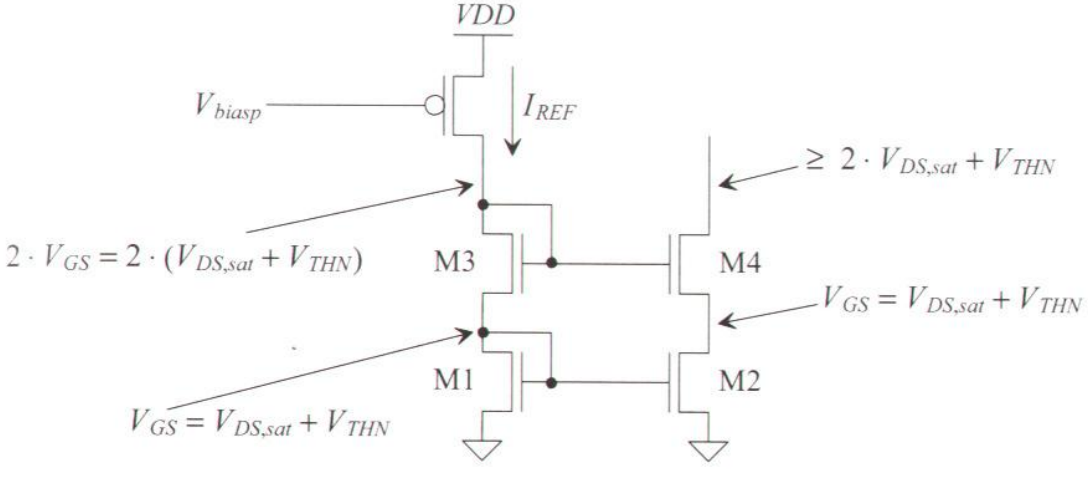
  Connecting node N to the gate M3, we have

  $\boldsymbol{V_{GS0} + V_X = V_{GS3} + V_Y}$

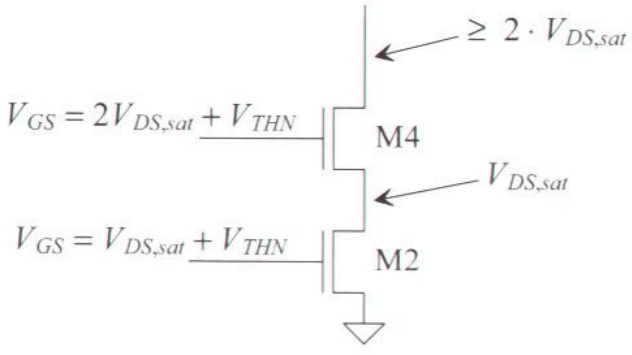
- Thus, if $\dfrac{(W/L)_3}{(W/L)_0} = \dfrac{(W/L)_2}{(W/L)_1}$, then $\boldsymbol{V_{GS0} = V_{GS3}}$ and $\boldsymbol{V_X = V_Y}$

The concept of shielding property of cascodes **The high output impedance of a cascode means that if the** output node voltage **is changed by** ΔV**, the resulting change at the** source of the cascode **is much less. In a sense, the cascode transistor "shileds" the input device from voltage variations at the output.**

The shielding property of cascodes diminishes if the cascode device enters the triode region.

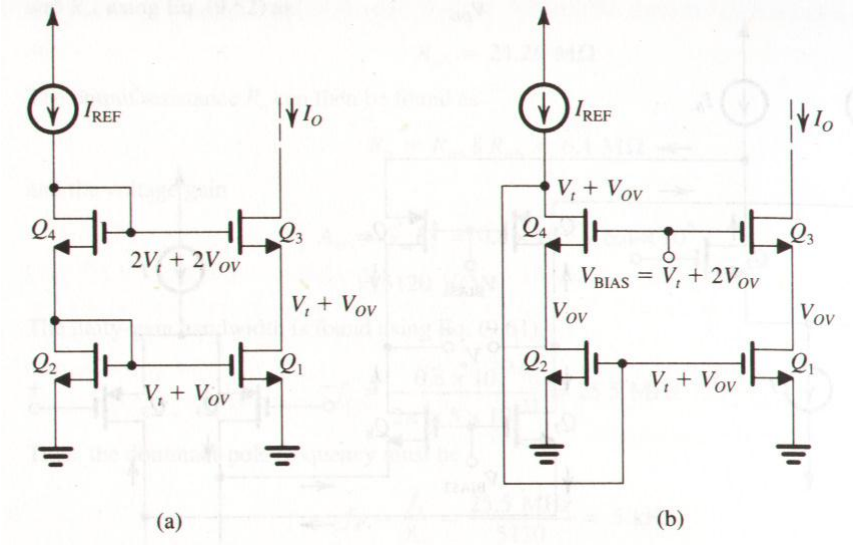


(a) Regular cascode structure



(b) Low-voltage (aka wide-swing) structure

Wide Swing Cascode Current Mirror



Because the voltage at the gate of Q_3 is $\mathbf{2V_t+2V_{OV}}$, the minimum voltage permitted at the output (while Q_3 remains saturated) is $\mathbf{V_t+2V_{OV}}$, hence the extra $\mathbf{V_t}$.

$V_{DS} = V_{GS} - V_t$
$V_D - V_S = V_G - V_S - V_t$
$V_D = V_G - V_t$
$V_D = (2V_t+2V_{OV}) - V_t$
$\mathbf{V_D = V_t+2V_{OV}}$

Also observe that Q_1 is operating with a drain-to-source voltage V_t+V_{OV} , which is V_t volts greater than it needs to operate in saturation.

To permit the output voltage at the drain of Q_3 to swing as low as $2V_{OV}$, we must lower the voltage at the gate of Q_3 from $2V_t+2V_{OV}$ to V_t+2V_{OV} .

$V_{DS} = V_{GS} - V_t$
$V_D - V_S = V_G - V_S - V_t$
$V_D = V_G - V_t$
$V_D = (V_t+2V_{OV}) - V_t$
$\mathbf{V_D = 2V_{OV}}$

However we can no longer connect the gate of Q_2 to its drain. Rather, it is connected to the drain of Q_4. This establishes a voltage of V_t+V_{OV} at the drain of Q_4 which is **sufficient to operate Q_4 is saturation**.

$V_{GS4} = V_{G4} - V_{S4} = (V_t+2V_{OV}) - V_{OV} = V_t+V_{OV}$
$V_{GS4} = V_{DS4}$, this makes **Q_4 always in saturation**

In order to get $2V_{OD}$, the drain current must be $4I_{in}$.

$$V_{REF} = 2V_{OD} + V_{tn}$$
$$= 2(V_{GS} - V_{tn}) + V_{tn}$$
$$= 2V_{GS} - V_{tn}$$

$$I_{REF} = \frac{k_n}{2}\frac{W}{L}(V_{GS} - V_{tn})^2$$

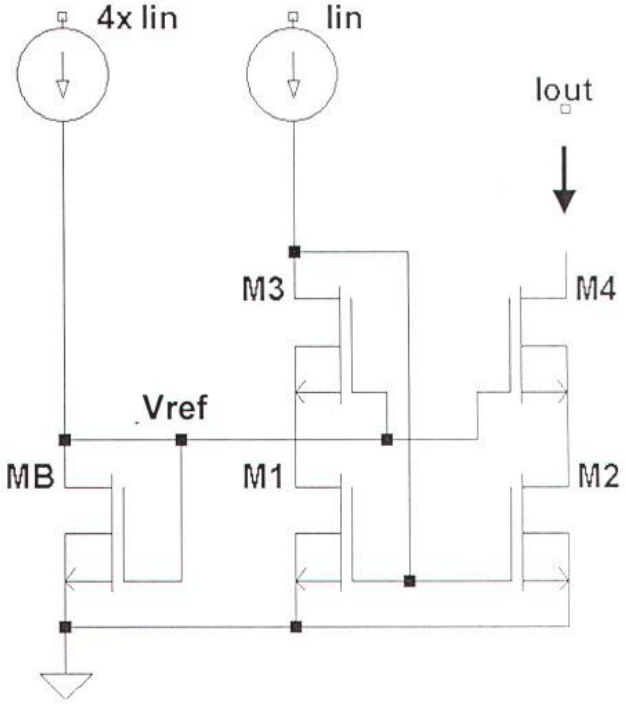
$$I_{REF} = \frac{k_n}{2}\frac{W}{L}(V_{REF} - V_{tn})^2$$
$$= \frac{k_n}{2}\frac{W}{L}(2V_{GS} - V_{tn} - V_{tn})^2$$
$$= \frac{k_n}{2}\frac{W}{L}(2V_{GS} - 2V_{tn})^2$$
$$= \frac{k_n}{2}\frac{W}{L}4(V_{GS} - V_{tn})^2$$
$$I_{REF} = 4I_{in}$$

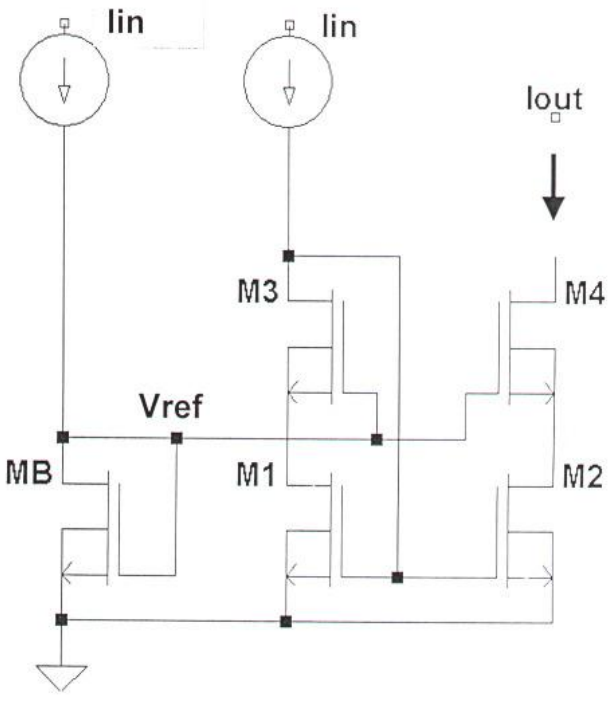
$$V_{OD} = \sqrt{\frac{4I_{in}}{k_n/2} \times \frac{L}{W}}$$ OR

$$V_{OD} = \sqrt{\frac{I_{in}}{k_n/2} \times \frac{4L}{W}}$$

Thus,

$$L_{MB} = 4L_1$$



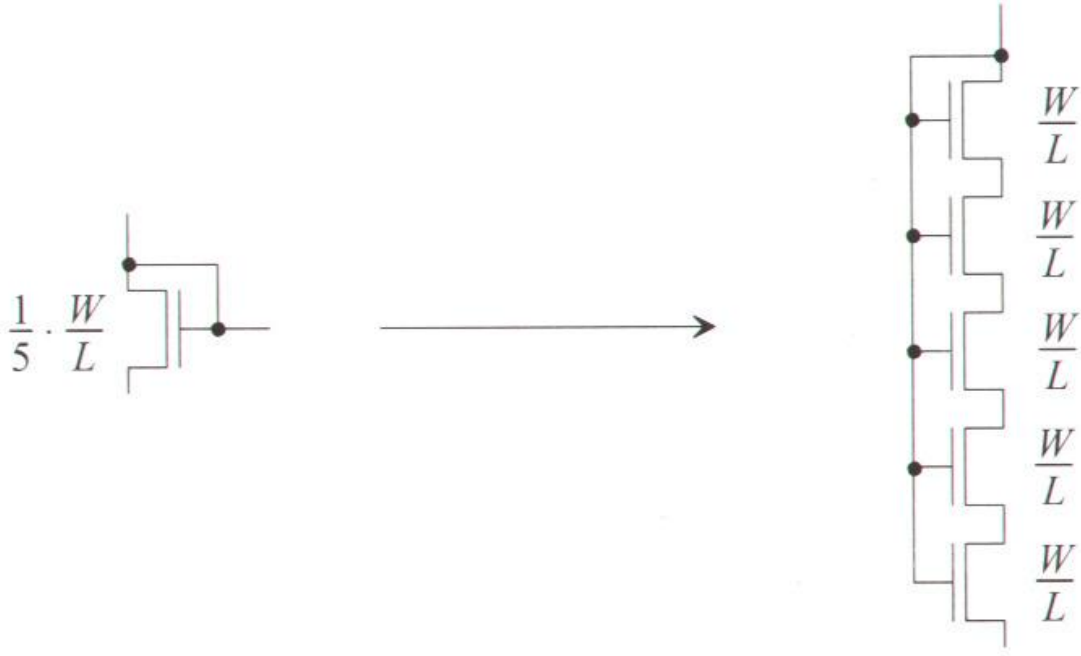


Layout Concerns

- When we go to layout the long L device, we might simply layout a single MOSFET with the appropriate length. However, the threshold voltage can vary significantly with the length of the device.

- Solution for this problem is by connecting MOSFETs in series with the same widths and their gates tied together behave like a single MOSFET with the sum of the individual MOSFET's lenghts.

- Because each device is identical. Changes in the threshold voltage shouldn't affect the biasing circuit.



UNIT - 3

MOS AND BICMOS CIRCUIT DESIGN PROCESSES: Mass layers, strick diagrams, design, symbolic diagrams

8 Hours

3.1 Introduction

In this chapter, the basic mask layout design guidelines for CMOS logic gates will be presented. The design of physical layout is very tightly linked to overall circuit performance (area, speed, power dissipation) since the physical structure directly determines the transconductances of the transistors, the parasitic capacitances and resistances, and obviously, the silicon area which is used for a certain function. On the other hand, the detailed mask layout of logic gates requires a very intensive and time-consuming design effort, which is justifiable only in special circumstances where the area and/or the performance of the circuit must be optimized under very tight constraints. Therefore, automated layout generation (e.g., standard cells + computer-aided placement and routing) is typically preferred for the design of most digital VLSI circuits. In order to judge the physical constraints and limitations, however, the VLSI designer must also have a good understanding of the physical mask layout process.

Mask layout drawings must strictly conform to a set of layout design rules as described in Chapter 2, therefore, we will start this chapter with the review of a complete design rule set. The design of a simple CMOS inverter will be presented step-by-step, in order to show the influence of various design rules on the mask structure and on the dimensions. Also, we will introduce the concept of stick diagrams, which can be used very effectively to simplify the overall topology of layout in the early design phases. With the help of stick diagrams, the designer can have a good understanding of the topological constraints, and quickly test several possibilities for the optimum layout without actually drawing a complete mask diagram.

The physical (mask layout) design of CMOS logic gates is an iterative process which starts with the circuit topology (to realize the desired logic function) and the initial sizing of the transistors (to realize the desired performance specifications). At this point, the designer can only estimate the total parasitic load at the output node, based on the fan-out, the number of devices, and the expected length of the interconnection lines. If the logic gate contains more than 4-6 transistors, the topological graph representation and the Euler-path method allow the designer to determine the optimum ordering of the transistors. A simple stick diagram layout can now be drawn, showing the locations of the transistors, the local interconnections between the transistors and the locations of the contacts.

After a topologically feasible layout is found, the mask layers are drawn (using a layout editor tool) according to the layout design rules. This procedure may require several small iterations in order to accommodate all design rules, but the basic topology should not change very significantly. Following the final DRC (Design Rule Check), a circuit extraction procedure is performed on the finished layout to determine the actual transistor sizes, and more importantly, the parasitic capacitances at each node. The result of the extraction step is usually a detailed

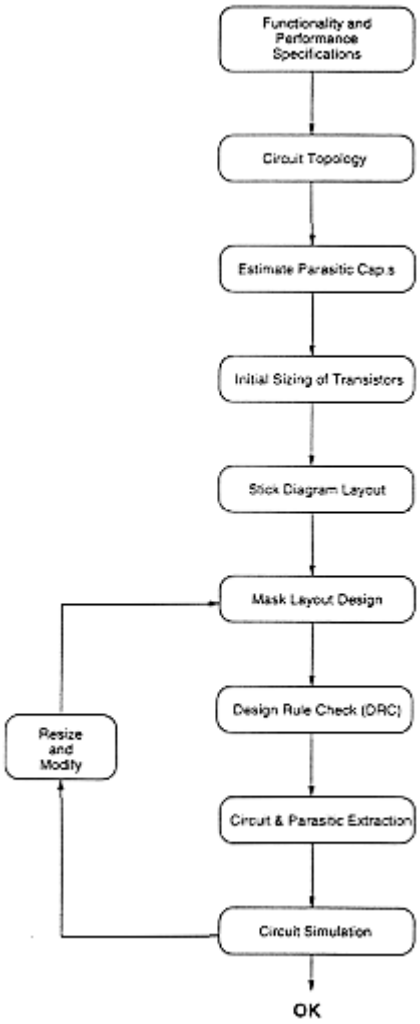


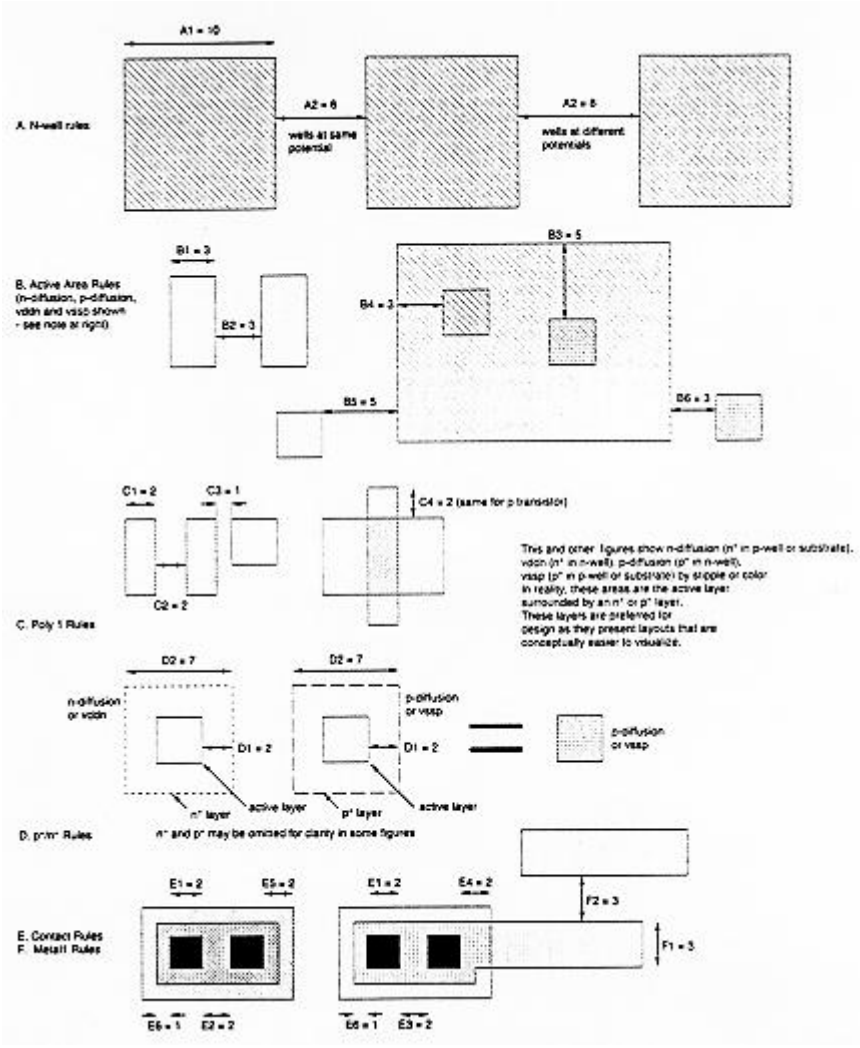
Figure-3.1: The typical design flow for the production of a mask layout.

SPICE input file, which is automatically generated by the extraction tool. Now, the actual performance of the circuit can be determined by performing a SPICE simulation, using the extracted net-list. If the simulated circuit performance (e.g., transient response times or power dissipation) do not match the desired specifications, the layout must be modified and the whole process must be repeated. The layout modifications are usually concentrated on the (W/L) ratios of the transistors (transistor re-sizing), since the width-to-length ratios of the transistors determine the device transconductance and the parasitic source/drain capacitances. The designer may also decide to change parts or all of the circuit topology in order to reduce the parasitics. The flow diagram of this iterative process is shown in Fig. 3.1.

3.2 CMOS Layout Design Rules

As already discussed in Chapter 2, each mask layout design must conform to a set of layout design rules, which dictate the geometrical constraints imposed upon the mask layers by the technology and by the fabrication process. The layout designer must follow these rules in order to guarantee a certain yield for the finished product, i.e., a certain ratio of acceptable chips out of a fabrication batch. A design which violates some of the layout design rules may still result in a functional chip, but the yield is expected to be lower because of random process variations.

The design rules below are given in terms of scaleable lambda-rules. Note that while the concept of scaleable design rules is very convenient for defining a technology-independent mask layout and for memorizing the basic constraints, most of the rules do not scale linearly, especially for sub-micron technologies. This fact is illustrated in the right column, where a representative rule set is given in real micron dimensions. A simple comparison with the lambda- based rules shows that there are significant differences. Therefore, lambda-based design rules are simply not useful for sub-micron CMOS technologies.



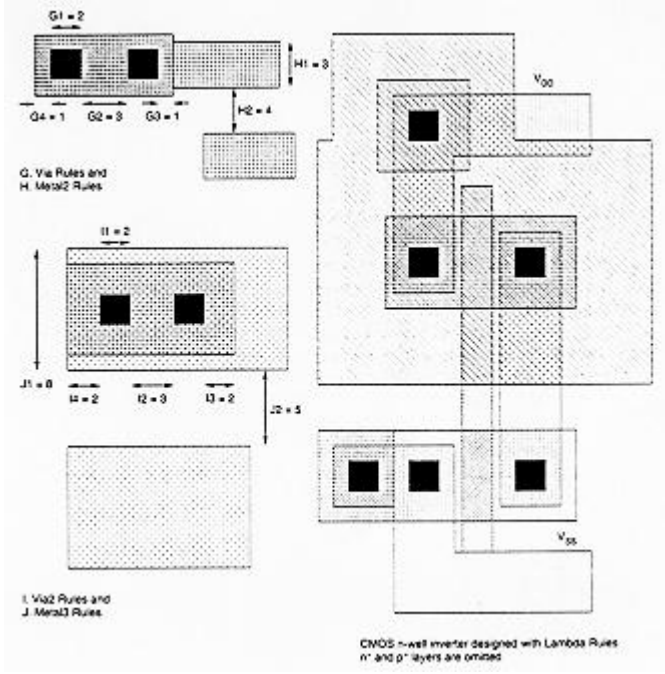


Figure-3.2: Illustration of CMOS layout design rules.

3.3 CMOS Inverter Layout Design

In the following, the mask layout design of a CMOS inverter will be examined step-by-step. The circuit consists of one nMOS and one pMOS transistor, therefore, one would assume that the layout topology is relatively simple. Yet, we will see that there exist quite a number of different design possibilities even for this very simple circuit.

First, we need to create the individual transistors according to the design rules. Assume that we attempt to design the inverter with minimum-size transistors. The width of the active area is then determined by the minimum diffusion contact size (which is necessary for source and drain connections) and the minimum separation from diffusion contact to both active area edges. The width of the polysilicon line over the active area (which is the gate of the transistor) is typically taken as the minimum poly width (Fig. 3.3). Then, the overall length of the active area is simply determined by the following sum: (minimum poly width) + 2 x (minimum poly-to- contact spacing) + 2 x (minimum spacing from contact to active area edge). The pMOS transistor must be placed in an n-well region, and the minimum size of the n- well is dictated by the pMOS active area and the minimum n-well overlap over n+. The distance between the nMOS and the pMOS transistor is determined by the minimum separation between the n+ active area and the n-well (Fig. 3.4). The polysilicon gates of the nMOS and the pMOS transistors are usually aligned. The final step in the mask layout is the local interconnections in metal, for the output node and for the VDD and GND contacts (Fig. 3.5). Notice that in order to be biased properly, the n-well region must also have a VDD contact.

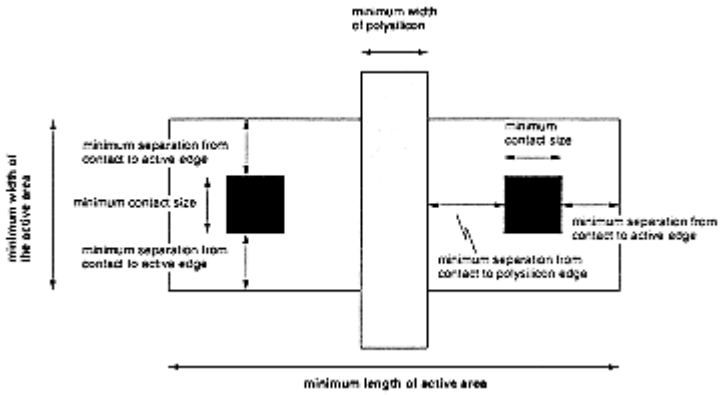


Figure-3.3: Design rule constraints which determine the dimensions of a minimum-size transistor.

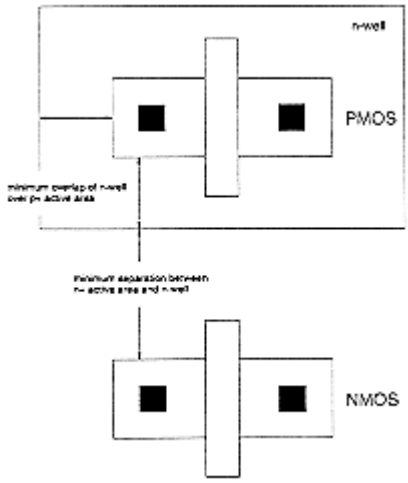


Figure-3.4: Placement of one nMOS and one pMOS transistor.

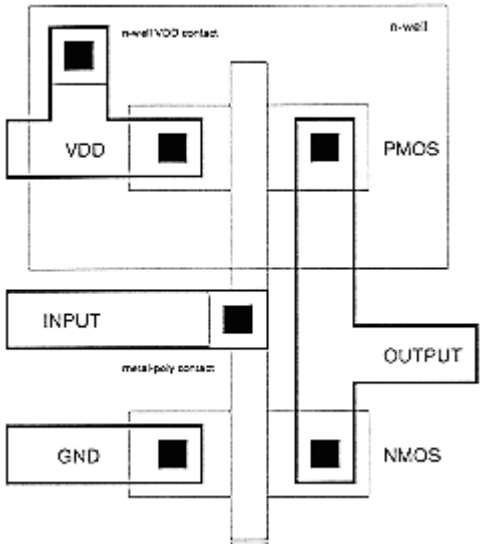


Figure-3.5: Complete mask layout of the CMOS inverter.

The inital phase of layout design can be simplified significantly by the use of stick diagrams - or so-called symbolic layouts. Here, the detailed layout design rules are simply neglected and the main features (active areas, polysilicon lines, metal lines) are represented by constant width rectangles or simple sticks. The purpose of the stick diagram is to provide the designer a good understanding of the topological constraints, and to quickly test several possibilities for the optimum layout without actually drawing a complete mask diagram. In the following, we will examine a series of stick diagrams which show different layout options for the CMOS inverter circuit.

The first two stick diagram layouts shown in Fig. 3.6 are the two most basic inverter configurations, with different alignments of the transistors. In some cases, other signals must be routed over the inverter. For instance, if one or two metal lines have to be passed through the middle of the cell from left to right, horizontal metal straps can be used to access the drain terminals of the transistors, which in turn connect to a vertical Metal-2 line. Metal-1 can now be used to route the signals passing through the inverter. Alternatively, the diffusion areas of both transistors may be used for extending the power and ground connections. This makes the inverter transistors transparent to horizontal metal lines which may pass over.

The addition of a second metal layer allows more interconnect freedom. The second- level metal can be used for power and ground supply lines, or alternatively, it may be used to vertically strap the input and the output signals. The final layout example in Fig. 3.6 shows one possibility of using a third metal layer, which is utilized for routing three signals on top.

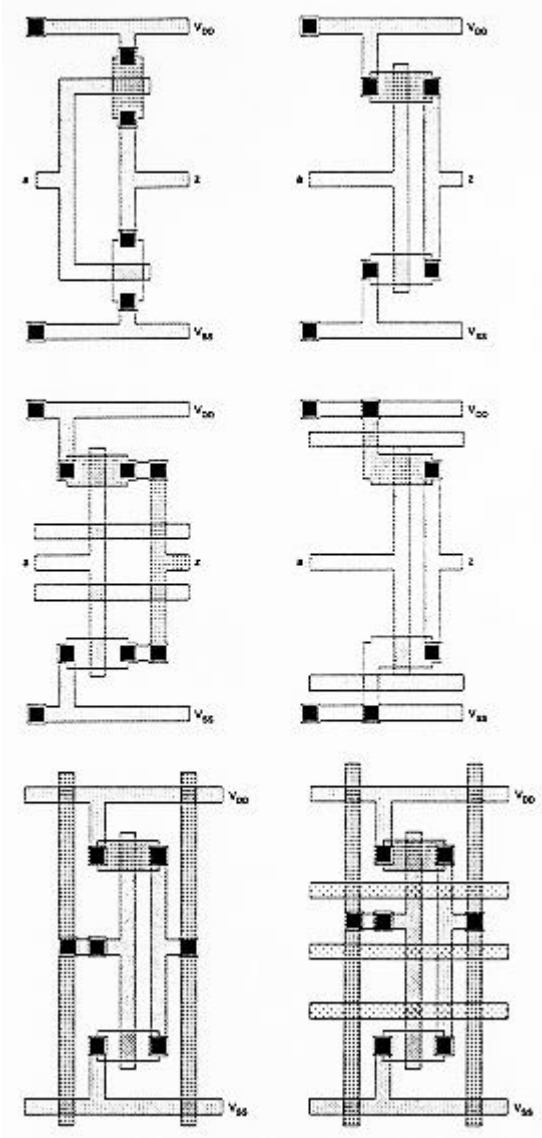


Figure-3.6: Stick diagrams showing various CMOS inverter layout options.

3.4 Layout of CMOS NAND and NOR Gates

The mask layout designs of CMOS NAND and NOR gates follow the general principles examined earlier for the CMOS inverter layout. Figure 3.7 shows the sample layouts of a two- input NOR gate and a two-input NAND gate, using single-layer polysilicon and single-layer metal. Here, the p-type diffusion area for the pMOS transistors and the n-type diffusion area for the nMOS transistors are aligned in parallel to allow simple routing of the gate signals with two parallel polysilicon lines running vertically. Also notice that the two mask layouts show a very strong symmetry, due to the fact that the NAND and the NOR gate are have a symmetrical circuit topology. Finally, Figs 3.8 and 3.9 show the major steps of the mask layout design for both gates, starting from the stick diagram and progressively defining the mask layers.

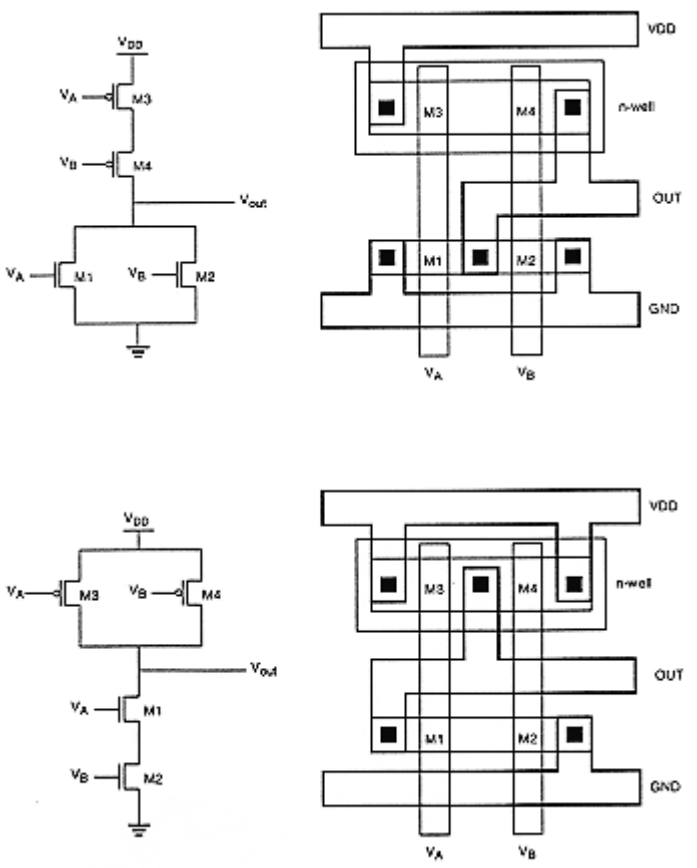


Figure-3.7: Sample layouts of a CMOS NOR2 gate and a CMOS NAND2 gate.

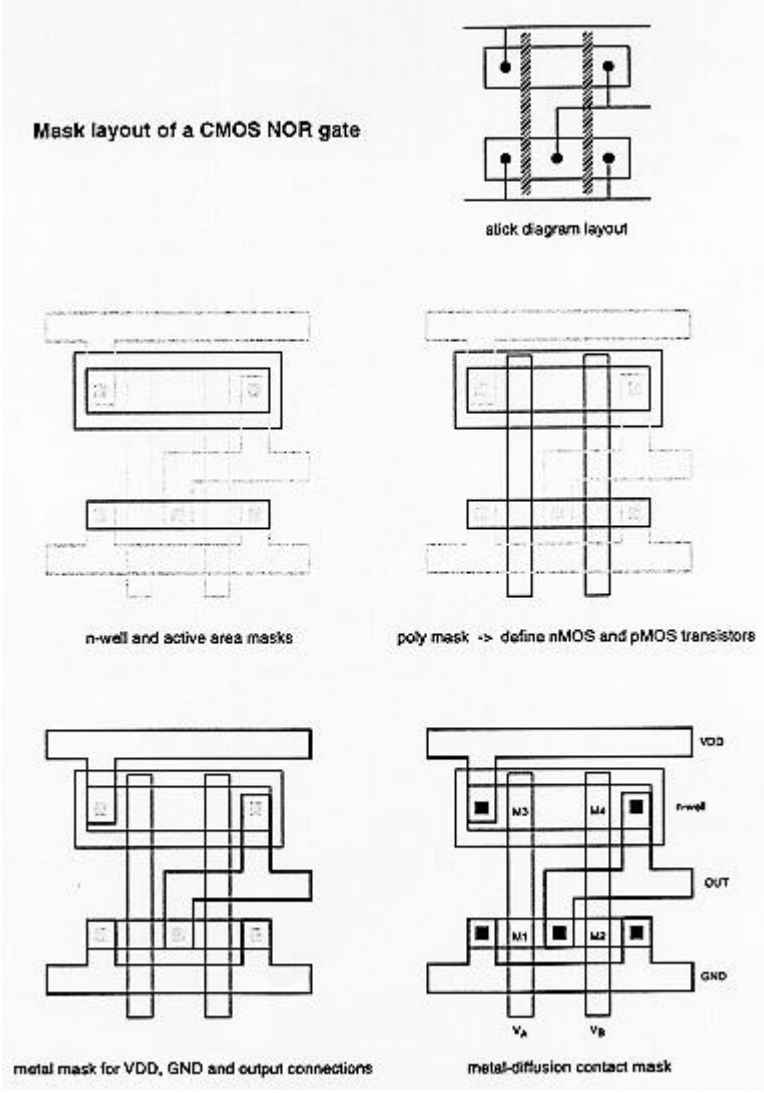


Figure-3.8: Major steps required for generating the mask layout of a CMOS NOR2 gate.

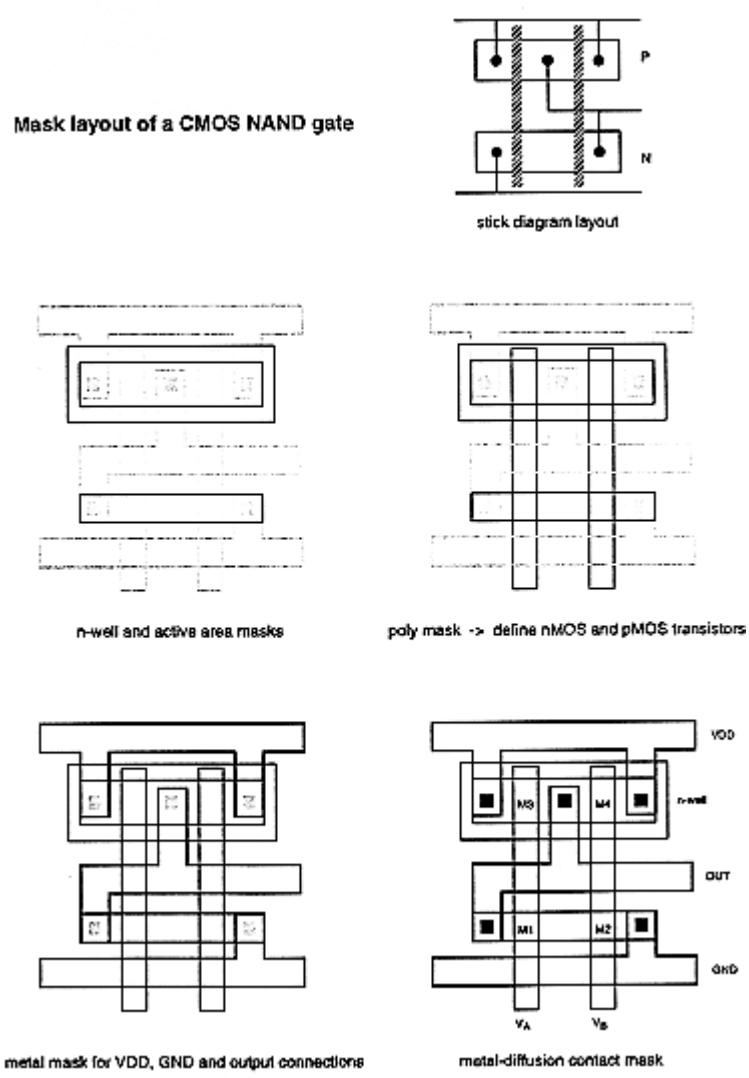


Figure-3.9: Major steps required for generating the mask layout of a CMOS NAND2 gate.

3.5 Complex CMOS Logic Gates

The realization of complex Boolean functions (which may include several input variables and several product terms) typically requires a series-parallel network of nMOS transistors which constitute the so-called pull-down net, and a corresponding dual network of pMOS transistors which constitute the pull-up net. Figure 3.10 shows the circuit diagram and the corresponding network graphs of a complex CMOS logic gate. Once the network topology of the nMOS pull- down network is known, the pull-up network of pMOS transistors can easily be constructed by using the dual-graph concept.

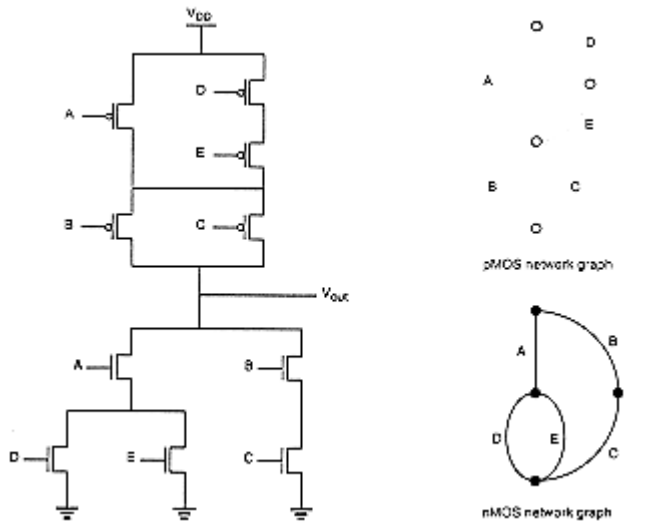


Figure-3.10: A complex CMOS logic gate realizing a Boolean function with 5 input variables.

Now, we will investigate the problem of constructing a minimum-area layout for the complex CMOS logic gate. Figure 3.11 shows the stick-diagram layout of a "first-attempt", using an arbitrary ordering of the polysilicon gate columns. Note that in this case, the separation between the polysilicon columns must be sufficiently wide to allow for two metal-diffusion contacts on both sides and one diffusion-diffusion separation. This certainly consumes a considerable amount of extra silicon area.

If we can minimize the number of active-area breaks both for the nMOS and for the pMOS transistors, the separation between the polysilicon gate columns can be made smaller. This, in turn, will reduce the overall horizontal dimension and the overall circuit layout area. The number of active-area breaks can be minimized by changing the ordering of the polysilicon columns, i.e., by changing the ordering of the transistors.

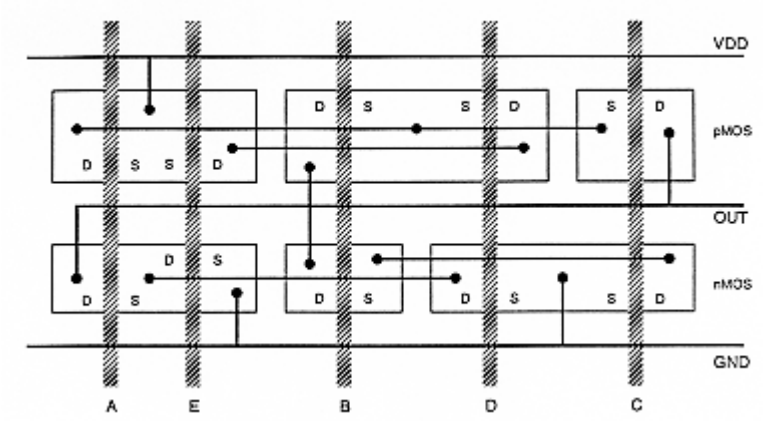


Figure-3.11: Stick diagram layout of the complex CMOS logic gate, with an arbitrary ordering of the polysilicon gate columns.

A simple method for finding the optimum gate ordering is the Euler-path method: Simply find a Euler path in the pull-down network graph and a Euler path in the pull-up network graph with the identical ordering of input labels, i.e., find a common Euler path for both graphs. The Euler path is defined as an uninterrupted path that traverses each edge (branch) of the graph exactly once. Figure 3.12 shows the construction of a common Euler path for both graphs in our example.

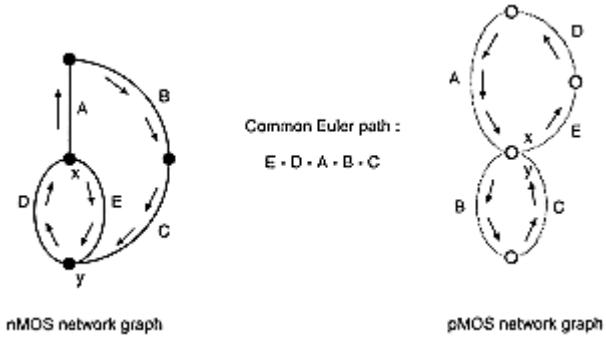


Figure-3.12: Finding a common Euler path in both graphs for the pull-down and pull-up net provides a gate ordering that minimizes the number of active-area breaks. In both cases, the Euler path starts at (x) and ends at (y).

It is seen that there is a common sequence (E-D-A-B-C) in both graphs. The polysilicon gate columns can be arranged according to this sequence, which results in uninterrupted active areas for nMOS as well as for pMOS transistors. The stick diagram of the new layout is shown in Fig. 3.13. In this case, the separation between two neighboring poly columns must allow only for one metal-diffusion contact. The advantages of this new layout are more compact (smaller) layout area, simple routing of signals, and correspondingly, smaller parasitic capacitance.

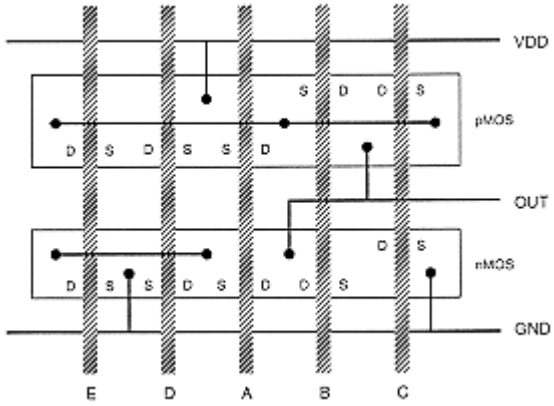


Figure-3.13: Optimized stick diagram layout of the complex CMOS logic gate.

It may not always be possible to construct a complete Euler path both in the pull-down and in the pull-up network. In that case, the best strategy is to find sub-Euler-paths in both graphs, which should be as

long as possible. This approach attempts to maximize the number of transistors which can be placed in a single, uninterrupted active area.

Finally, Fig. 3.14 shows the circuit diagram of a CMOS one-bit full adder. The circuit has three inputs, and two outputs, sum and carry_out. The corresponding mask layout of this circuit is given in Fig. 3.15. All input and output signals have been arranged in vertical polysilicon columns. Notice that both the sum-circuit and the carry-circuit have been realized using one uninterrupted active area each.

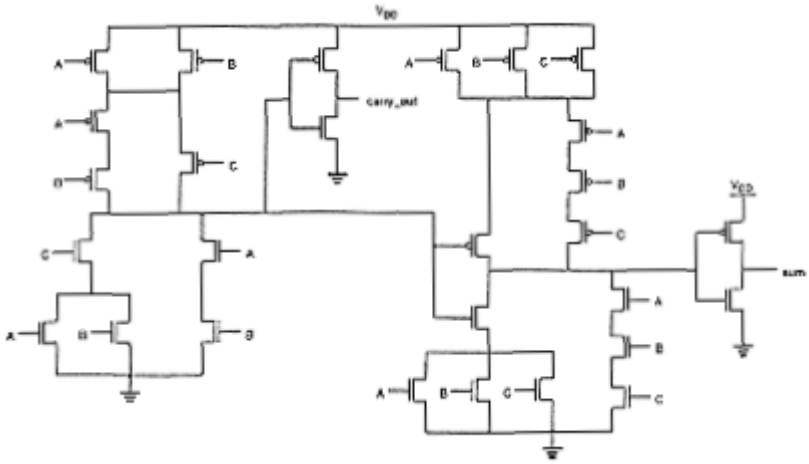


Figure-3.14: Circuit diagram of the CMOS one-bit full adder.

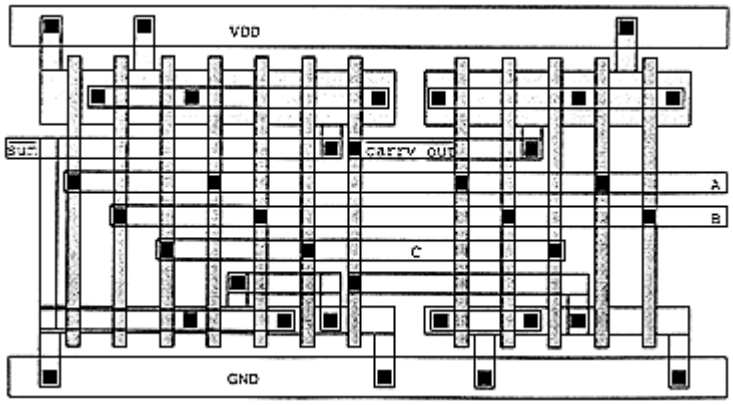


Figure-3.15: Mask layout of the CMOS full adder circuit..

The preceding lectures have already given you the information of the different layers, their representation (colour,hatching)etc. When the devices are represented using these layers, we call it physical design. The design is carried out using the design tool, which requires to follow certain rules. Physical structure is required to study the impact of moving from circuit to layout. When we draw the layout from the schematic, we are taking the first step towards the physical design.

Physical design is an important step towards fabrication. Layout is representation of a schematic into layered diagram. This diagram reveals the different layers like ndiff, polysilicon etc that go into formation of the device.

At every stage of the physical design simulations are carried out to verify whether the design is as per requirement. Soon after the layout design the DRC check is used to verify minimum dimensions and spacing of the layers. Once the layout is done, a layout versus schematic check carried out before proceeding further. There are different tools available for drawing the layout and simulating it.

The simplest way to begin a layout representation is to draw the stick diagram. But as the complexity increases it is not possible to draw the stick diagrams. For beginners it easy to draw the stick diagram and then proceed with the layout for the basic digital gates . We will have a look at some of the things we should know before starting the layout.

In the schematic representation lines drawn between device terminals represent interconnections and any no planar situation can be handled by crossing over. But in layout designs a little more concern about the physical interconnection of different layers. By simply drawing one layer above the other it not possible to make interconnections, because of the different characters of each layer. Contacts have to be made whenever such interconnection is required. The power and the ground connections are made using the metal and the common gate connection using the polysilicon. The metal and the diffusion layers are connected using contacts. The substrate contacts are made for same source and substrate voltage. which are not implied in the schematic. These layouts are governed by DRC's and have to be atleast of the minimum size depending on the technology used . The crossing over of layers is another aspect which is of concern and is addressed next.

1.Poly crossing diffusion makes a transistor
2.Metal of the same kind crossing causes a short.

3.Poly crossing a metal causes no interaction unless a contact is made.

Different design tricks need to be used to avoid unknown creations. Like a combination of metal1 and metal2 can be used to avoid short. Usually metat2 is used for the global vdd and vss lines and metal1 for local connections.

SCHEMATIC AND LAYOUT OF BASIC GATES

1.CMOS INVERTERNOT GATE) SCHEMATIC

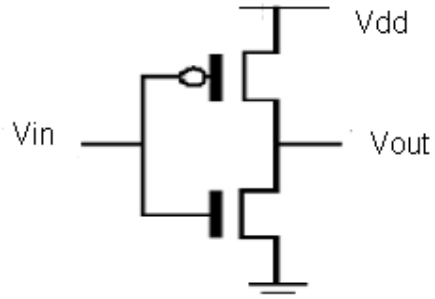


Figure 1 Inverter

TOWARDS THE LAYOUT

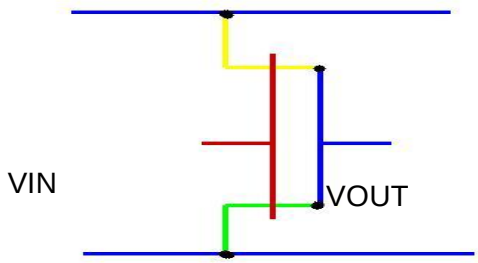


Figure2: Stick diagram of inverter

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

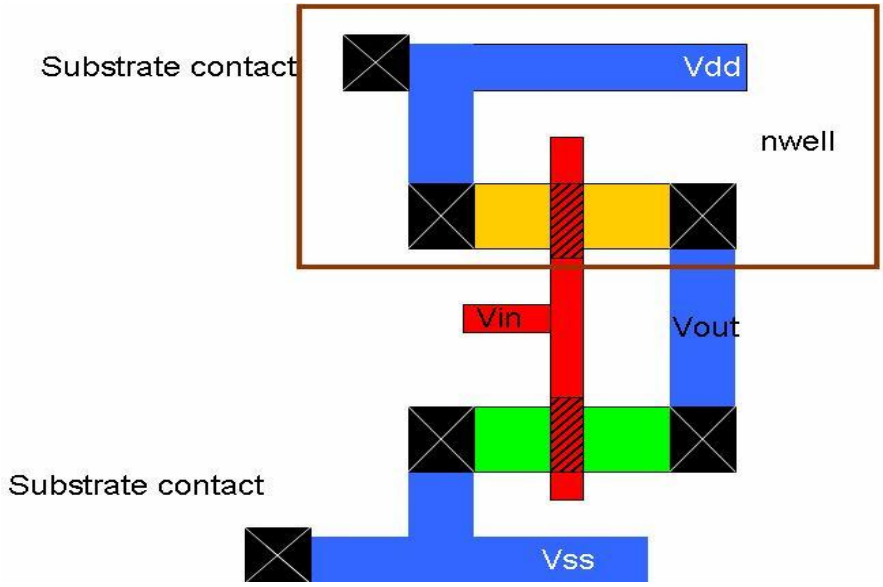


Figure 3: Layout of an inverter

The layout shown above is that of a CMOS inverter. It consists of a pdiff (yellow colour) forming the pmos at the junction of the diffusion and the polysilicon(red colour)shown hatched ndiff(green) forming the nmos(area hatched).The different layers drawn are checked for their dimensions using the DRC rule check of the tool used for drawing. Only after the DRC(design rule check) is passed the design can proceed further. Further the design undergoes Layout Vs Schematic checks and finally the parasitic can be extracted.

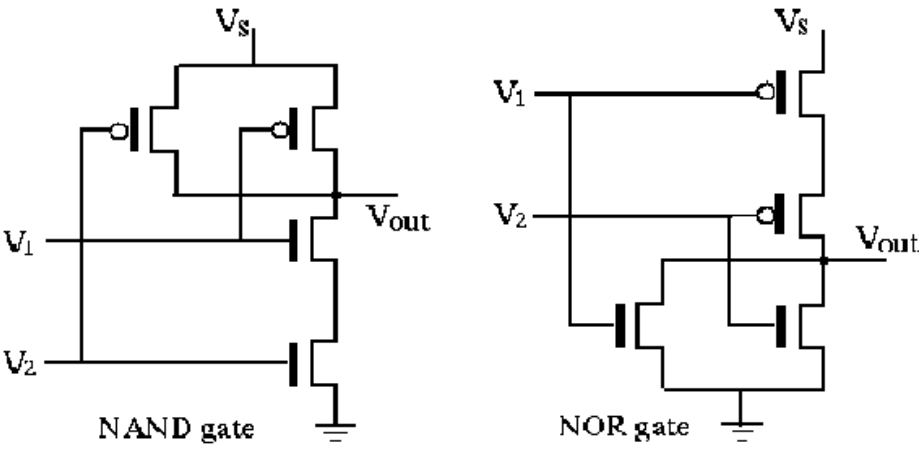


Figure 4:Schematic diagrams of nand and nor gate

We can seen that the nand gate consists of two pmos in parallel which forms the pull up logic and two nmos in series forming the pull down logic.It is the complementary for the nor gate. We get inverted logic from cmos structures. The series and parallel connections are for getting the right logic output. The pull up and the pull down devices must be placed to get high and low outsputs when required.

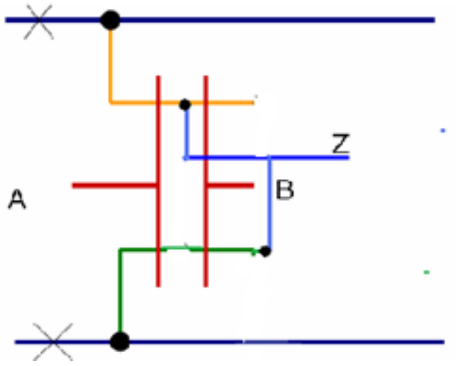


Figure 5: Stick diagram of nand gate

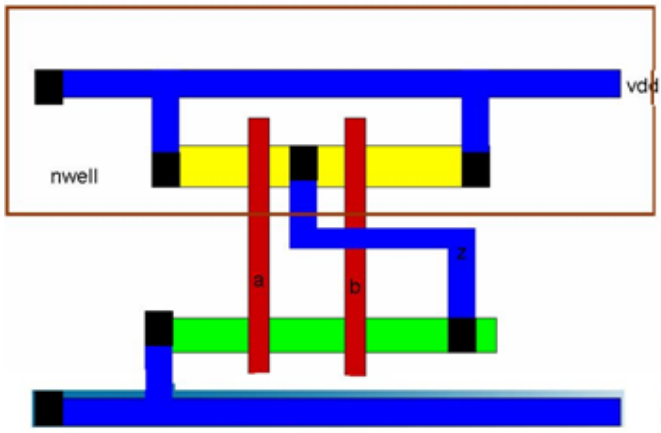


Figure 6: Layout of a nand gate

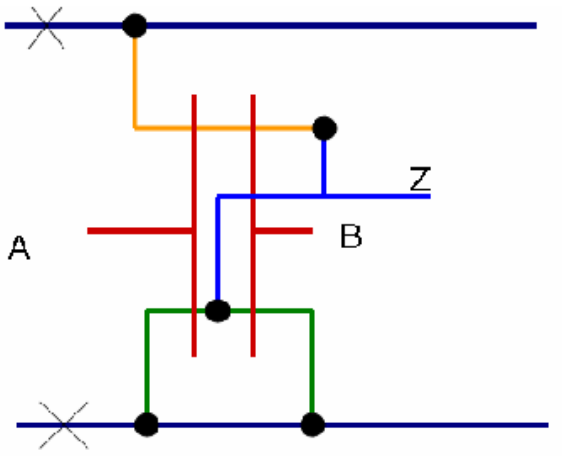


Figure 7:Stick diagram of nor gate

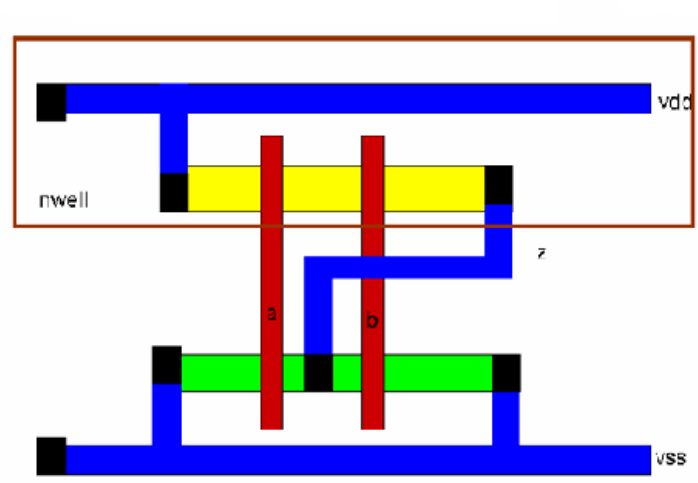


Figure 8: Layout of nor gate

TRANSMISSION GATE

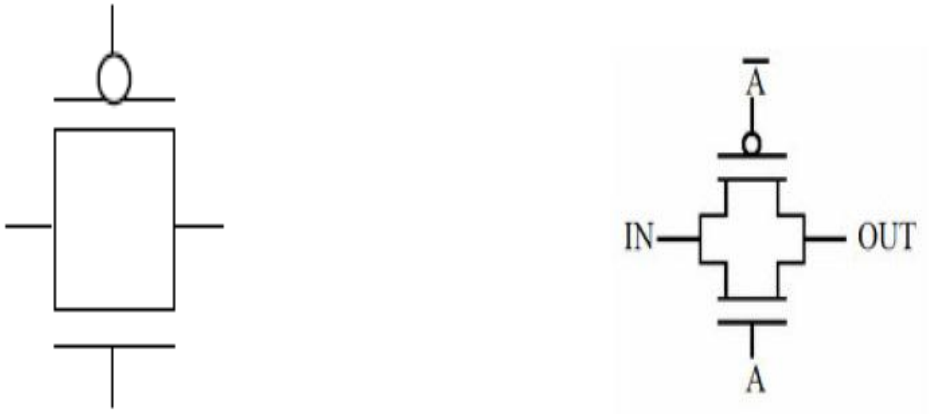


Figure 9 :Symbol and schematic of transmission gate

Layout considerations of transmission gate. It consist of drains and the sources of the P&N devices paralleled. Transmission gate can replace the pass transistors and has the advantage of giving both a good one and a good zero.

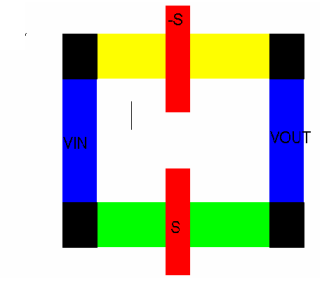


Figure 10: Layout of transmissuion gate

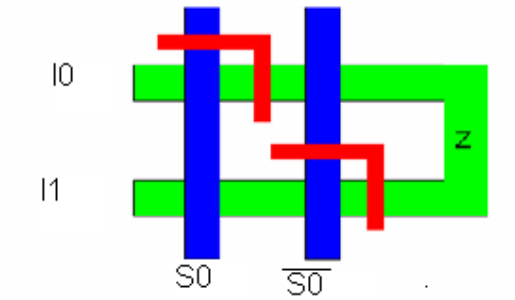


Figure 11:TG with nmos switches

CMOS STANDARD CELL DESIGN

Geometric regularity is very important to maintain some common electrical characteristics between the cells in the library. The common physical limitation is to fix the height and vary the width according to the required function. The Wp and Wn are fixed considering power dissipation, propagation delay, area and noise immunity. The best thing to do is to fix a required objective function and then fix Wn and Wp to obtain the required objective

Usually in CMOS Wn is made equal to Wp . In the process of designing these gates techniques may be employed to automatically generate the gates of common size. Later optimization can be carried out to achieve a specific feature. Gate array layout and sea of gate layout are constructed using the above techniques. The gate arrays may be customized by having routing channels in between array of gates. The gate array and the sea of gates have some special layout considerations. *The gate arrays* use fixed image of the under layers i.e the diffusion and poly are fixed and metal are programmable. The wiring layers are discretionary and providing the personalization of the array. The rows of transistors are fixed and the routing channels are provided in between them. Hence the design issues involves size of transistors, connectivity of poly and the number of routing channels required.

Sea of gates in this style continuous rows of n and p diffusion run across the master chip and are arranged without regard to the routing channel. Finally the routing is done across unused transistors saving space.

GENERAL LAYOUT GUIDELINES

1.The electrical gate design must be completed by checking the following a.Right power

and ground supplies

b.Noise at the gate input

c.Faulty connections and transistors d.Improper

ratios

c.Incorrect clocking and charge sharing

2.VDD and the VSS lines run at the top and the bottom of the design

3.Vertical poysilicon for each gate input

4.Order polysilicon gate signals for maximal connection between transistors

5.The connectivity requires to place nmos close to VSS and pmos close to VDD

6.Connection to complete the logic must be made using poly,metal and even metal2

The design must always proceeds towards optimization. Here optimization is at transistor level rather then gate level. Since the density of transistors is large ,we could obtain smaller and faster layout by designing logic blocks of 1000 transistors instead of considering a single at a time and then putting them together. Density improvement can also be made by considering optimization of the other factors in the layout

The factors are

1.Efficient routing space usage. They can be placed over the cells or even in multiple layers.

2.Source drain connections must be merged better.

3.White (blank) spaces must be minimum

4.The devices must be of optimum sizes.

5.Transperent routing can be provided for cell to cell interconnection, this reduces global wiring problems

LAYOUT OPTIMIZATION FOR PERFORMANCE

1.Vary the size of the transistor according to its position in series. The transistor closest to the output is the smallest. The transistor nearest to the VSS line is the largest. This helps in increasing the performance by 30 %. A three input nand gate with the varying size is shown next.

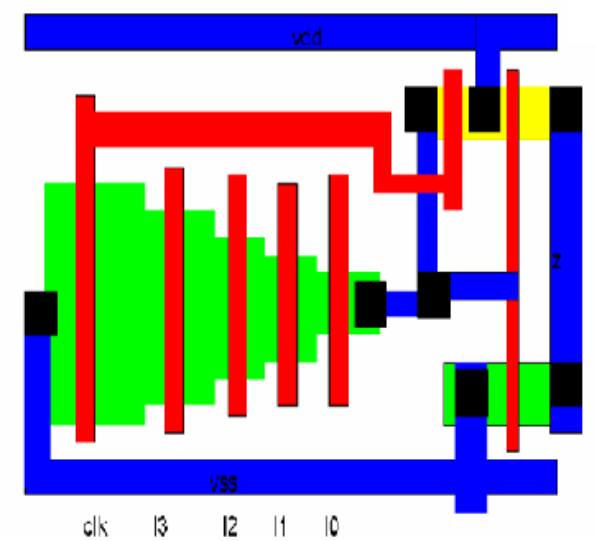


Figure 12 :Layout optimization with varying diffusion areas

2. Less optimized gates could occur even in the case of parallel connected transistors.This is usually seen in parallel inverters, nor & nand.When drains are

connected in parallel ,we must try and reduce the number of drains in parallel ie wherever possible we must try and connect drains in series at least at the output.This arrangement could reduce the capacitance at the output enabling good voltage levels. One example is as shown next.

Figure 13 Layout of nor gate showing series and parallel drains

UNIT - 4

BASIC CIRCUIT CONCEPTS: Sheet resistance, capacitance layer inverter delays, wiring capacitance, choice of layers.

BASIC CIRCUIT DESIGN CONCEPTS

INTRODUCTION

We have already seen that MOS structures are formed by the super imposition of a number conducting ,insulating and transistor forming material. Now each of these layers have their own characteristics like capacitance and resistances. These fundamental components are required to estimate the performance of the system. These layers also have inductance characteristics that are important for I/O behaviour but are usually neglected for on chip devices.

The issues of prominence are

1.Resistance, capacitance and inductance calculations.

2.Delay estimations

3.Determination of conductor size for power and clock distribution

4.Power consumption

5.Charge sharing

6.Design margin

7.Reliabiltiy

8.Effects and extent of scaling

RESISTANCE ESTIMATION

The concept of sheet resistance is being used to know the resistive behavior of the layers that go into formation of the MOS device. Let us consider a uniform slab of conducting material of the following characteristics .

Resistivity- ρ Width -

W Thickness -

t

Length between faces – L as shown next

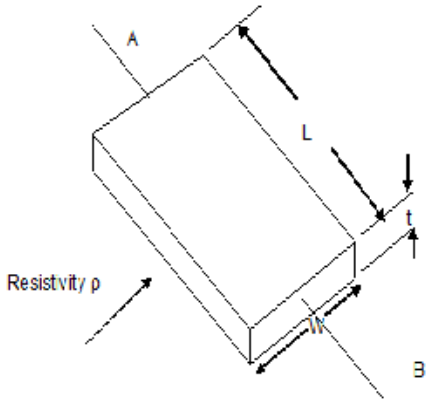


Figure 24:A slab of semiconductor

We know that the resistance is given by RAB= ρL/A Ω. The area of the slab considered above is given by A=Wt. There fore RAB= ρL/Wt Ω. If the slab is considered as a square then L=W. therefore RAB=ρ/t which is called as sheet resistance represented by Rs.The unit of sheet resistance is **ohm per square.** It is to be noted that Rs is independent of the area of the slab. Hence we can conclude that a

1um per side square has the same resistance as that of 1cm per side square of the same material.

The resistance of the different materials that go into making of the MOS device depend on the resistivity and the thickness of the material. For a diffusion layer the depth defines the thickness and the impurity defines the resistivity. The table of values for a 5u technology is listed below.5u technology means minimum line width is 5u and ƛ= 2.5u.The diffusion mentioned in the table is n diffusion, p diffusion values are

2.5 times of that of n. The table of standard sheet resistance value follows.

Layer	Rs per square
Metal	0.03
Diffusion n(for 2.5 times the n)	10 to 50
Silicide	2 to 4
Polysilicon	15 to 100
N transistor gate	10^4
P transistor gate	$2.5x\ 10^4$

SHEET RESISTANCE OF MOS TRANSISTORS

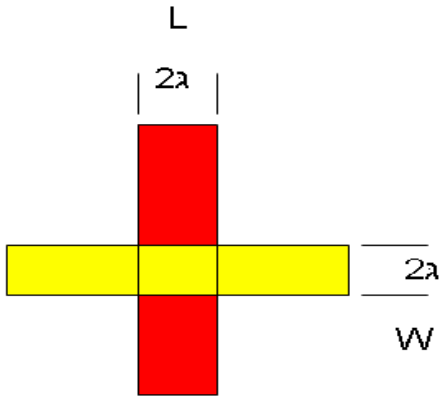


Figure 25 Min sized inverter

The N transistor above is formed by a 2λ wide poly and n diffusion. The L/W ratio is

1. Hence the transistor is a square, therefore the resistance R is 1sqxRs ohm/sq i.e. R=1x104. If L/W ratio is 4 then R = $4x10^4$. If it is a P transistor then for L/W =1,the value of R is $2.5x10^4$.

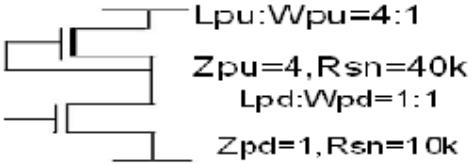


Figure 26 Nmos depletion inverter

Pull up to pull down ratio = 4.In this case when the nmos is on, both the devices are on simultaneously, Hence there is an on resistance Ron = 40+10 =50k. It is this resistance that leads the static power consumption which is the disadvantage of nmos depletion mode devices

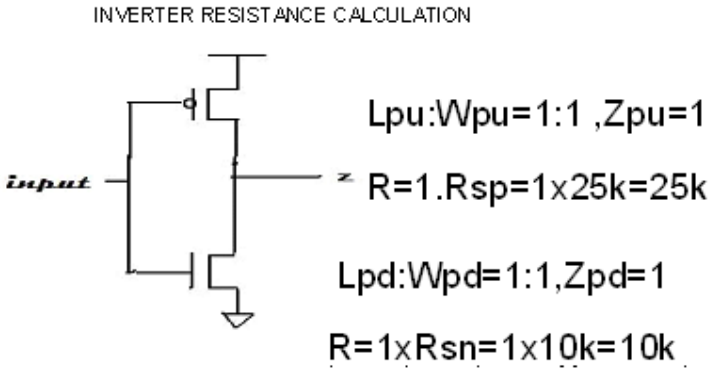


Figure 27: Cmos inverter

Since both the devices are not on simultaneously there is no static power dissipation

The resistance of non rectangular shapes is a little tedious to estimate. Hence it is easier to convert the irregular shape into regular rectangular or square blocks and then estimate the resistance. For example

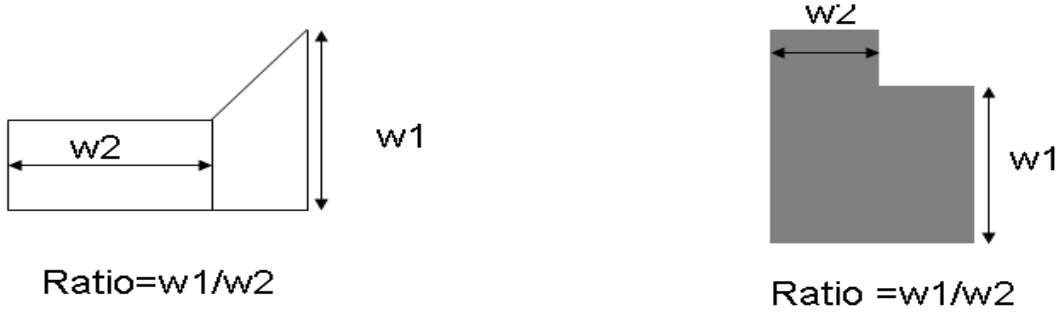


Figure 28:Irregular rectangular shapes

CONTACT AND VIA RESISTANCE

The contacts and the vias also have resistances that depend on the contacted materials and the area of contact. As the contact sizes are reduced for scaling ,the associated resistance increases. The resistances are reduced by making ohmic contacts which are also called loss less contacts. Currently the values of resistances vary from .25ohms to a few tens of ohms.

SILICIDES

The connecting lines that run from one circuit to the other have to be optimized. For this reason the width is reduced considerably. With the reduction is width the sheet resistance increases, increasing the RC delay component. With poly silicon the sheet resistance values vary from 15 to 100 ohm. This actually effects the extent of scaling down process. Polysilicon is being replaced with silicide. Silicide is obtained by depositing metal on polysilicon and then sintering it. Silicides give a sheet resistance of 2 to 4 ohm. The reduced sheet resistance makes silicides a very attractive replacement for poly silicon. But the extra processing steps is an offset to the advantage.

A Problem

A particular layer of MOS circuit has a resistivity ρ of 1 ohm –cm. The section is 55um long,5um wide and 1 um thick. Calculate the resistance and also find Rs

R= RsxL/W, Rs= ρ/t Rs=1x10-2/1x10-6=10^4ohm R= 104x55x10-6/5x106=110k

CAPACITANCE ESTIMATION

Parasitics capacitances are associated with the MOS device due to different layers that go into its formation. Interconnection capacitance can also be formed by the metal, diffusion and polysilicon (these are often called as runners) in addition with the transistor and conductor resistance. All these capacitances actually define the switching speed of the MOS device.

Understanding the source of parasitics and their variation becomes a very essential part of the design specially when system performance is measured in terms of the speed. The various capacitances that are associated with the CMOS device are

1.Gate capacitance - due to other inputs connected to output of the device

2.Diffusion capacitance - Drain regions connected to the output

3.Routing capacitance- due to connections between output and other inputs

The fabrication process illustrates that the conducting layers are apparently seperated from the substrate and other layers by the insulating layer leading to the formation of parallel capacitors. Since the silicon dioxide is the insulator knowing its thickness we can calculate the capacitance

$$C = \frac{\varepsilon_o \varepsilon_{ins} A}{D} \quad \text{farad}$$

εεo= permittivity of free space- 8.854x1014f/cm

εins= relative permitivity of sio2=4.0
D= thickness of the dioxide in cm
A = area of the plate in cm2

The gate to channel capacitance formed due to the sio2 separation is the most profound of the mentioned three types. It is directly connected to the input and the output. The other capacitance like the metal, poly can be evaluated against the substrate. The gate capacitance is therefore standardized so as to enable to move from one technology to the other conveniently.

The standard unit is denoted by ¤ Cg. It represents the capacitance between gate to channel with W=L=min feature size. Here is a figure showing the different capacitances that add up to give the total gate capacitance

Cgd, Cgs = gate to channel capacitance lumped at the source and drain

Csb, Cdb = source and drain diffusion capacitance to substrate

Cgb = gate to bulk capacitance

Total gate capacitance Cg = Cgd+Cgs+Cgb

Since the standard gate capacitance has been defined, the other capacitances like polysilicon, metal, diffusion can be expressed in terms of the same standard units so that the total capacitance can be obtained by simply adding all the values. In order to express in standard values the following steps must be followed

1. Calculate the areas of area under consideration relative to that of standard gate i.e.$4\lambda^2$. (standard gate varies according to the technology)
2. Multiply the obtained area by relative capacitance values tabulated .
3. This gives the value of the capacitance in the standard unit of capacitance ¤ Cg.

Table 1:Relative value of Cg

layer	**Relative value for5u technology**
Gate to channel	1
Diffusion	0.25
Poly to sub	0.1
M1 to sub	0.075
M2 to sub	0.05

M2 to M1	0.1
M2 to poly	0.075

For a 5u technology the area of the minimum sized transistor is 5uX5u=25um2 ie λ=2.5u, hence,area of minimum sized transistor in lambda is 2 λX 2 λ = 4 λ2.Therefore for 2u or 1.2u or any other technology the area of a minimum sized transistor in lambda is 4 λ2. Lets solve a few problems to get to know the things better.

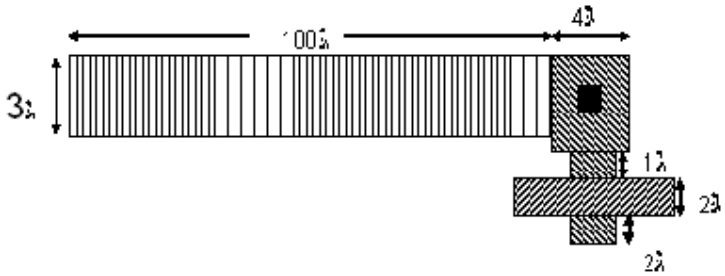


Figure 29 :Multilayered structure

The figure above shows the dimensions and the interaction of different layers, for evaluating the total capacitance resulting so.

Three capacitance to be evaluated metal Cm,polysilicon Cp and gate capacitance Cg

Area of metal = 100x3=300λ2

Relative area = 300/4=75

Cm=75Xrelative cap=75X0.075=5.625¤ Cg

Polysilicon capacitance Cp

Area of poly=(4x4+1x2+2X2)=22λ2

Relative area = 22λ2/4λ 2=5.5

Cp=5.5Xrelative cap=5.5x.1=0.55 ¤ Cg

Gate capacitance Cg= 1¤ Cg because it is a min size gate

Ct=Cm+Cp+Cg=5.625+0.55+1=7.2 ¤ Cg

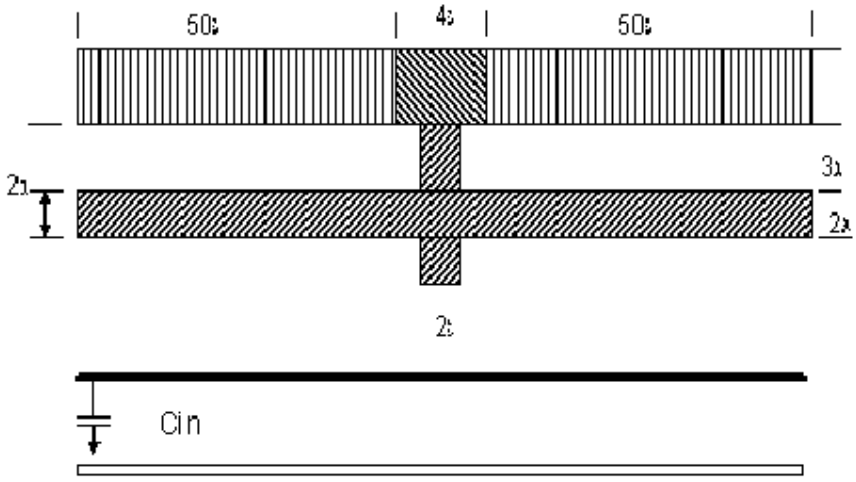


Figure 29:Mos structure

The input capacitance is made of three components metal capacitance Cm, poly capacitance Cp, gate capacitance Cg i.e Cin= Cm+Cg+Cp

Relative area of metal =(50x3)X2/4=300/4=75

Cm=75x0.075=5.625¤ Cg

Relative area of poly = (4x4+2x1+2x2)/4 =22/4 =5.5

Cp=5.5X0.1=0.55 ¤ Cg

Cg=1 ¤ Cg

Cin=7.175 ¤ Cg

Cout = Cd+Cperi. Assuming Cperi to be negligible. Cout = Cd.

Relative area of diffusion=51x2/4=102/4=25.5

Cd=25.5x0.25=6.25 ¤ Cg.

The relative values are for the 5um technology

DELAY

The concept of sheet resistance and standard unit capacitance can be used to calculate the delay. If we consider that a one feature size poly is charged by one feature size diffusion then the delay is Time constant 1Ɛ= Rs (n/p channel)x 1¤ Cg secs. This can be evaluated for any technology. The value of ¤ Cg will vary with different technologies because of the variation in the minimum feature size.

5u using n diffusion=104X0.01=0.1ns safe delay 0.03nsec

2um = 104x0.0032=0.064 nsecs safe delay 0.02nsec

1.2u= 104x0.0023 = 0.046nsecs safe delay =0.1nsec

These safe figures are essential in order to anticipate the output at the right time

INVERTER DELAYS

We have seen that the inverter is associated with pull up and pull down resistance values. Specially in nmos inverters. Hence the delay associated with the inverter will depend on whether it is being turned off or on. If we consider two inverters cascaded then the total delay will remain constant irrespective of the transitions. Nmos and Cmos inverter delays are shown next

NMOS INVERTER

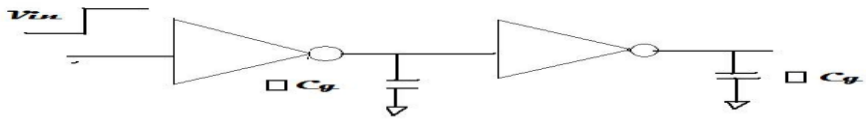


Figure 30: Cascaded nmos inverters

Let us consider the input to be high and hence the first inverter will pull it down. The pull down inverter is of minimum size nmos. Hence the delay is 1Ɛ. Second inverter will pull it up and it is 4 times larger, hence its delay is 4Ɛ.The total delay is 1Ɛ +4Ɛ= 5Ɛ. Hence for nmos the delay can be generalized as T=(1+Zpu/Zpd) Ɛ

CMOS INVERTER

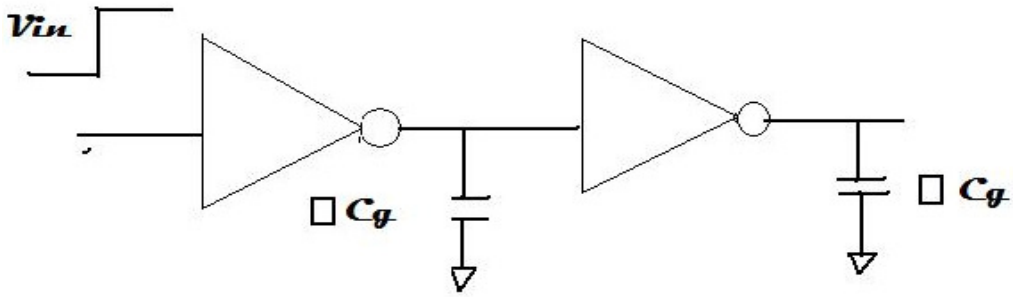


Figure 30 : Cascaded Cmos inverter

Let us consider the input to be high and hence the first inverter will pull it down. The nmos transistor has Rs = 10k and the capacitance is 2Cg. Hence the delay is 2Ɛ. Now the second inverter will pull it up, job done by the pmos. Pmos has sheet resistance of 25k i.e 2.5 times more, everything else remains same and hence delay is 5Ɛ. Total delay is 2Ɛ +5Ɛ = 7Ɛ . The capacitance here is double because the input is connected to the common poly, putting both the gate capacitance in parallel. The only factor to be considered is the resistance of the p gate which is increasing the delay. If want to reduce delay, we must reduce resistance. If we increase the width of p channel, resistance can be reduced but it increases the capacitance. Hence some trade off must be made to get the appropriate values.

FORMAL ESTIMATION OF DELAY

The inverter either charges or discharges the load capacitance CL. We could also estimate the delay by estimating the rise time and fall time theoritically.

Rise time estimation

Assuming that the p device is in saturation we have the current given by the equation
Idsp=ßp(Vgs-|Vtp|)2/2

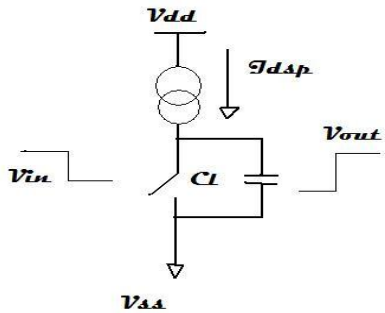


Figure 31 :Rise time estimation

The above current charges the capacitance and it has a constant value therefore the model can be written as shown in figure above. The output is the drop across the capacitance, given by

Vout =Idsp x t/CL

Substituting for Idsp we have Vout=ßp(Vgs-|Vtp|)2t/2CL. Therefore the equation for t=2CLVout/ßp(Vgs-|Vtp|).Let t=Ɛr and Vout=Vdd, therefore we have Ɛr = 2VddCL/ßp(Vgs-|Vtp|)2. If consider Vtp=0.2Vdd and Vgs=Vdd we have Ɛr = 3CL/ßpVdd

On similar basis the fall time can be also be written as Ɛf = 3CL/ßnVdd whose model can be written as shown next

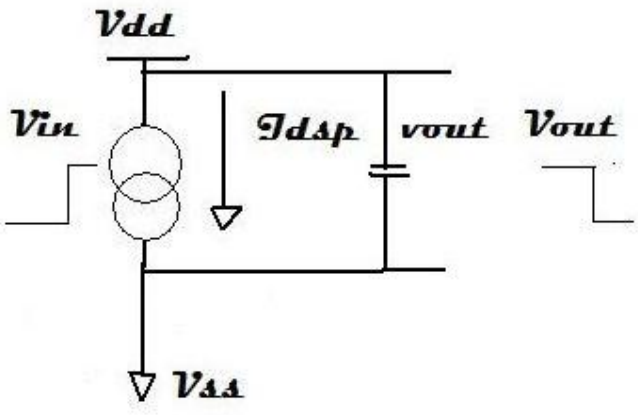


Figure 32 :Fall time estimation

DRIVING LARGE CAPACITIVE LOAD

The problem of driving large capacitive loads arises when signals must travel outside the chip. Usually it so happens that the capacitance outside the chip are higher. To reduce the delay these loads must be driven by low resistance. If we are using a cascade of inverter as drivers the pull and pull down resistances must be reduced. Low résistance means low L:W ratio. To reduce the ratio, W must be increased. Since L cannot be reduced to lesser than minimum we end up having a device which occupies a larger area. Larger area means the input capacitance increases and slows down the process more. The solution to this is to have N cascaded inverters with their sizes increasing, having the largest to drive the load capacitance. Therefore if we have

3 inverters,1st is smallest and third is biggest as shown next.

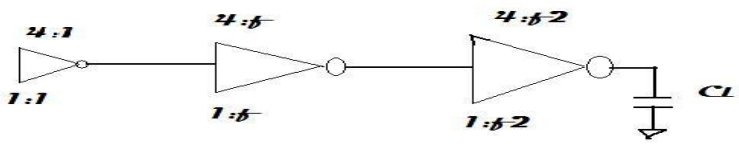


Figure 33:Cascaded inverters with varying widths

We see that the width is increasing by a factor of f towards the last stage. Now both f and N can be complementary. If f for each stage is large the number of stages N reduces but delay per stage increases. Therefore it becomes essential to optimize. Fix N and find the minimum value of f. For nmos inverters if the input transitions from 0 to 1 the delay is fε and if it transitions from 1 to 0 the delay is 4 fε. The delay for a nmos pair is 5 fε. For a cmos pair it will be 7fε

optimum value of f.

Assume y=CL/ ¤ Cg = fN, therefore choice of values of N and f are interdependent. We find the value of f to minimize the delay, from the equation of y we have ln(y)=Nln(f) i.e N=ln(y)/ln(f). If delay per stage is 5fε for nmos, then for even number of stages the total delay is N/2 5fε=2.5fε. For cmos total delay is N/2 7fε = 3.5fε

Hence delay ά Nft=ln(y)/ln(f)ft. Delay can be minimized if chose the value of f to be equal to e which is the base of natural logarithms. It means that each stage is 2.7wider than its predecessor. If f=e then N= ln(y).The total delay is then given by

1.For N=even

td=2.5Neε for nmos, td=3.5Neε for cmos

2.For N=odd

transition from 0 to 1 transition from1 to 0

td=[2.5(N-1)+1]eε nmos td=[2.5(N-1)+4]eε

td=[3.59N-1)+2]eε cmos td=[3.5(N-1)+5]eε

for example

For N=5 which is odd we can calculate the delay fro vin=1 as td=[2.5(5-1)+1]eε =11e ε

i.e. 1 +4+1+4+1 = 11eε

For vin =0 , td=[2.5(5-1)+4]eε = 14eε

4+1+4+1+4 = 14eε

SUPER BUFFER

The asymmetry of the inverters used to solve delay problems is clearly undesirable, this also leads to more delay problems, super buffer are a better solution. We have a inverting and non inverting variants of the super buffer. Such arrangements when used for 5u technology showed that they were capable of driving 2pf capacitance with 2nsec rise time.The figure shown next is the inverting variant.

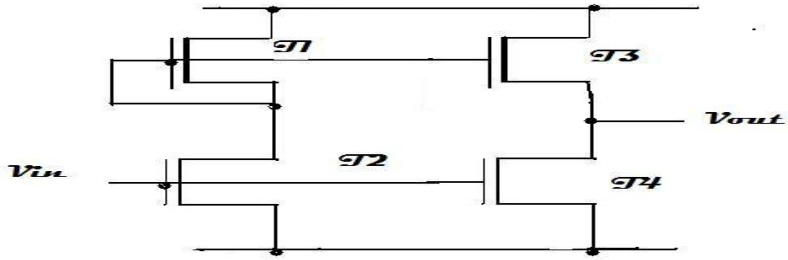


Figure 34:Inverterting buffer

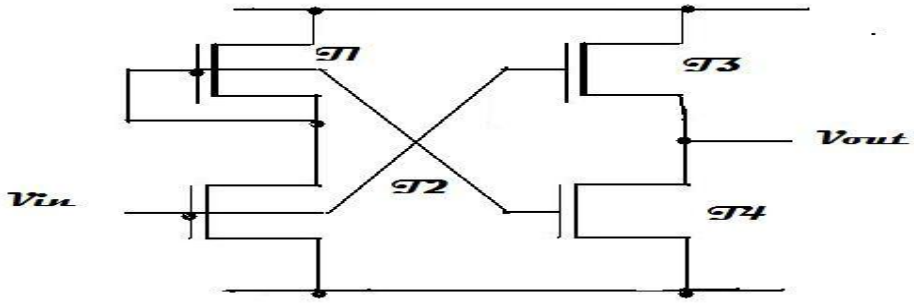
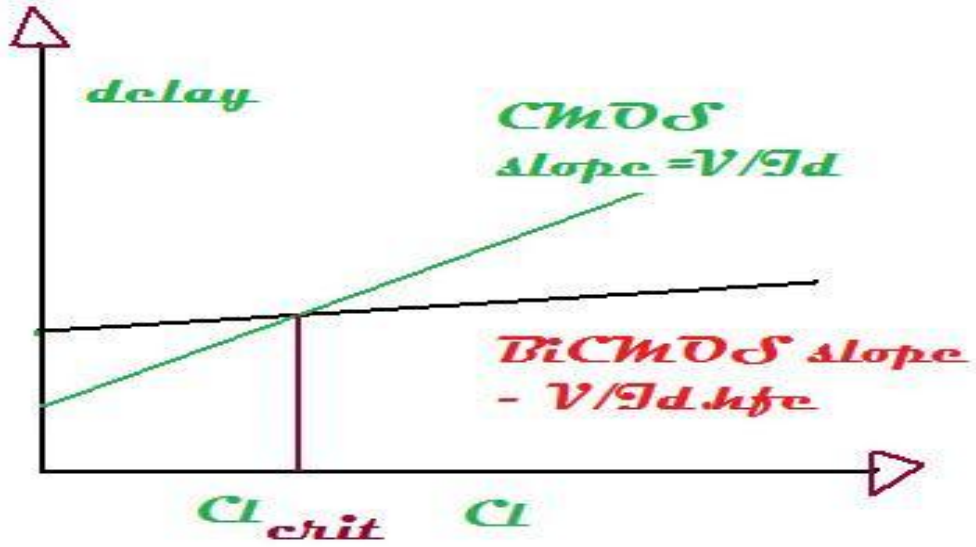


Figure 34:NonInverteing variant

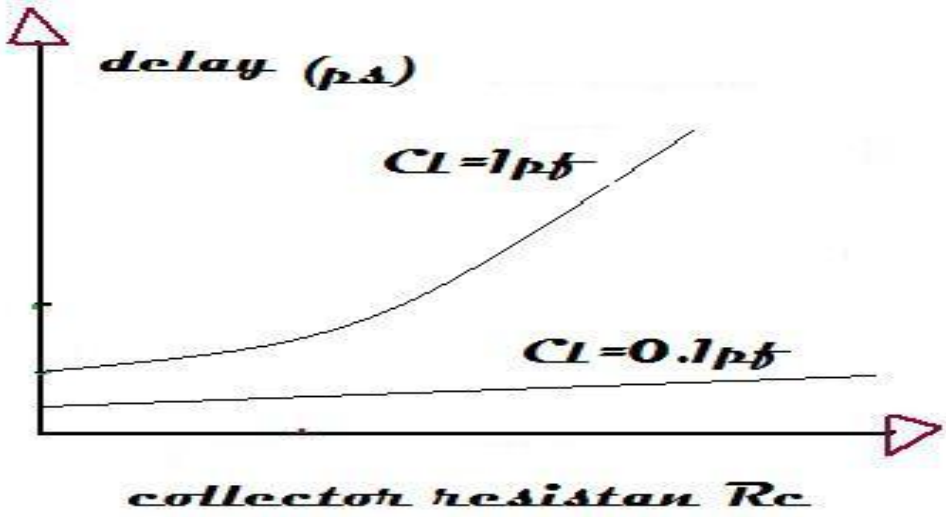
BICMOS DRIVERS

The availability of bipolar devices enables us to use these as the output stage of inverter or any logic. Bipolar devices have high Tran con--ductance and they are able switch large currents with smaller input voltage swings. The time required to change the out by an amount equal to the input is given by $\Delta t=CL/gm$, Where gm is the device trans conductance. Δt will be a very small value because of the high gm. The transistor delay consists of two components Tin and TL. Tin the time required to charge the base of the transistor which is large. TL is smaller because the time take to charge capacitor is less by hfe which is the transistor gain a comparative graph shown below.



Figure

The collector resistance is another parameter that contributes to the delay.The graph shown below shows that for smaller load capacitance, the delay is manageable but for large capacitance, as Rc increases the delay increase drastically.



Figure

By taking certain care during fabrication reasonably good bipolar devices can be produced with large hfe, gm ,ß and small Rc. Therefore bipolar devices used in buffers and logic circuits give the designers a lot of scpoe and freedom .This is coming without having to do any changes with the cmos circuit.

PROPAGATION DELAY

This is delay introduced when the logic signals have to pass through a chain of pass transistors. The transistors could pose a RC product delay and this increases drastically as the number of pass transistor in series increases.As seen from the figure the response at node V2 is given by CdV2/dt=(V1-V2)(V2-V3)/R For a long network we can

write RCdv/dt =dv2/dx2, i.e delay ά x2,

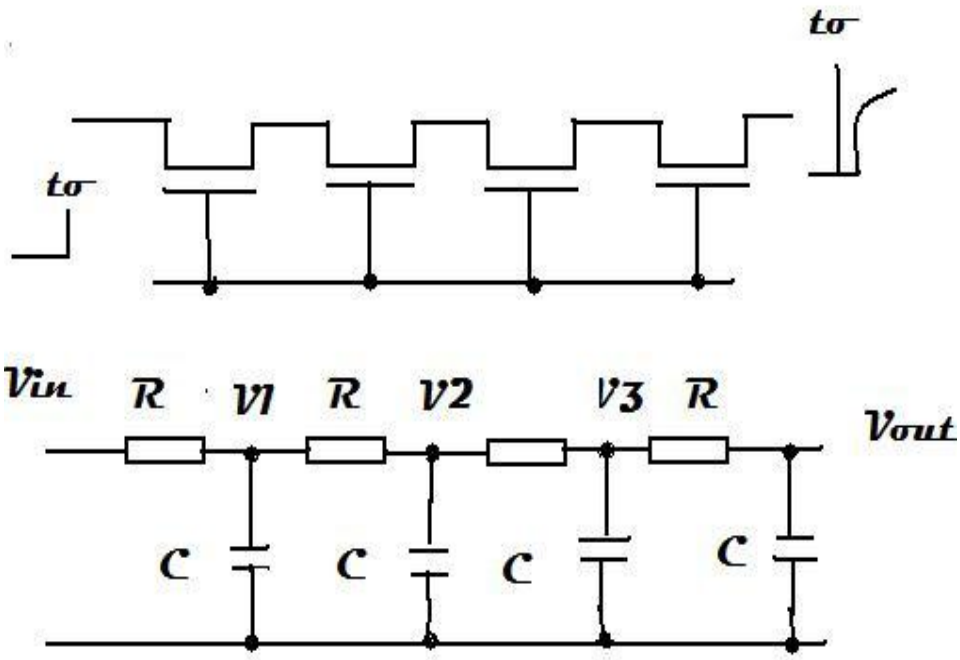


Figure 38

Lump all the R and C we have Rtotal=nrRs and C=nc□ Cg where and hence delay =n^2rcЄ. The increases by the square of the number, hence restrict the number of stages to maximum 4 and for longer ones introduce buffers in between.

DESIGN OF LONG POLYSILICONS

The following points must be considered before going in for long wire.

1.The designer is also discouraged from designing long diffusion lines also because the capacitance is much larger

2.When it inevitable and long poly lines have to used the best way to reduce delay is use buffers in between. Buffers also reduce the noise sensitivity

OTHER SOURCES OF CAPACITANCE Wiring capacitance

1.Fringing field

2.Interlayer capacitance

3.Peripheral capacitance

The capacitances together add upto as much capacitance as coming from the gate to source and hence the design must consider points to reduce them.The major of the wiring capacitance is coming from fringing field effects. Fringing capacitances is due to parallel fine metal lines running across the chip for power conection.The capacitance depends on the length l, thickness t and the distance d between the wire and thesubstrate. The

accurate prediction is required for performance estimation.Hence Cw=Carea+Cff.

Interlayer capacitance is seen when different layers cross each and hence it is neglected for simole calculations. Such capacitance can be easily estimated for regular structures and helps in modeling the circuit better.

Peripheral capacitance is seen at the junction of two devices. The source and the drain n regions form junctions with the pwell (substrate) and p diffusion form with adjacent nwells leading to these side wall (peripheral) capacitance

The capacitances are profound when the devices are shrunk in sizes and hence must be considered. Now the total diffusion capacitance is Ctotal = Carea + Cperi

In order to reduce the side wall effects, the designers consider to use isolation regions of alternate impurity.

CHOICE OF LAYERS

1.Vdd and Vss lines must be distributed on metal lines except for some exception

2.Long lengths of poly must be avoided because they have large Rs,it is not suitable for routing Vdd or Vss lines.

3.Since the resistance effects of the transistors are much larger, hence wiring effects due to voltage dividers are not that profound

Capacitance must be accurately calculated for fast signal lines usually those using high Rs material. Diffusion areas must be carefully handled because they have larger capacitance to substrate.

With all the above inputs it is better to model wires as small capacitors which will give electrical guidelines for communication circuits.

PROBLEMS

1.A particular section of the layout includes a 3λ wide metal path which crosses a 2 λ polysilicon path at right angles. Assuming that the layers are seperated by a 0.5 thick sio2,find the capacitance between the two.

Capacitance = Ɛ0 Ɛins A/D

Let the technology be 5um, λ=2.5um. Area = 7.5umX5um=37.5um C=4X8.854X10-12 x37.5/ 0.5 =2656pF The value of C in standard units is Relative area 6 λ2 /4 λ2 =1.5

C =1.5x0.075=0.1125¤ Cg

2 nd part of the problem

The polysilicon turns across a 4λ diffusion layer, find the gate to channel capacitance. Area = 2 λx 4λ =8 λ2 Relative area= 8 λ2 / 4 λ2 =2

Relative capacitance for 5u=1
Total gate capacitance = 2¤ Cg
Gate to channel capacitance>metal

2. The two nmos transistors are cascaded to drive a load capacitance of 16▯ Cg as shown in figure ,Calculate the pair delay. What are the ratios of each transistors. f stray and wiring capacitance is to be considered then each inverter will have an additional capacitance at the output of 4 ▯ Cg .Find the delay.

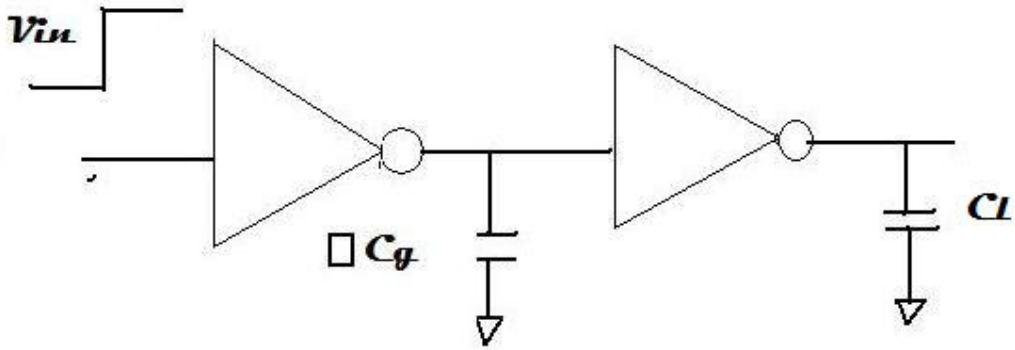


Figure 40

Lpu=16ƛ Wpu=2 ƛ Zpu=8

Lpd=2 ƛ Wpd=2 ƛ Zpd=1

Ratio of inverter 1 = 8:1

Lpu=2 ƛ Wpu=2 ƛ Zpu=1

Lpd =2 ƛ Wpd =8 ƛ Zpd=1/4

Ratio of inverter 2 = 1/1/4=4

Delay without strays

1Ɛ=Rsx1▯ Cg

Let the input transition from 1 to 0

Delay 1 = 8RsX▯ Cg=8Ɛ Delay 2=4Rs(▯ Cg +16 ▯ Cg)=68Ɛ Total delay = 76Ɛ

Delay with strays

Delay 1 = 8RsX(▯ Cg+ 4▯ Cg) = 40Ɛ Delay 2= 4RsX(▯ Cg+ 4▯ Cg +16 ▯ Cg)=84Ɛ

Total delay = 40+84=124Ɛ

If Ɛ = 0.1ns for 5u ie the delays are 7.6ns and 12.4ns

SCALING OF MOS DEVICES

The VLSI technology is in the process of evolution leading to reduction of the feature size and line widths. This process is called scaling down. The reduction in sizes has generally lead to better performance of the devices. There are certain limits on scaling and it becomes important to study the effect of scaling. The effect of scaling must be studied for certain parameters that effect the performance.

The parameters are as stated below

1.Minimum feature size

2.Number of gates on one chip

3.Power dissipation

4.Maximum operational frequency

5.Die size

6.Production cost .

These are also called as figures of merit

Many of the mentioned factors can be improved by shrinking the sizes of transistors, interconnects, separation between devices and also by adjusting the voltage and doping levels. Therefore it becomes essential for the designers to implement scaling and understand its effects on the performance

There are three types of scaling models used

1.Constant electric field scaling model

2.Constant voltage scaling model

3.Combined voltage and field model

The three models make use of two scaling factors 1/ß and 1/ά . 1/ß is chosen as the scaling factor for Vdd, gate oxide thickness D. 1/ ά is chosen as the scaling factor for all the linear dimensions like length, width etc. the figure next shows the dimensions and their scaling factors

The following are some simple derivations for scaling down the device parameters

1.Gate area Ag

Ag= L x W. Since L & W are scaled down by 1/ ά. Ag is scaled down by 1/ $ά^2$

2.Gate capacitance per unit area

Co=Ɛo/D, permittivity of sio2 cannot be scaled, hence Co can be scaled 1/1/ß=ß

3.Gate capacitance Cg

Cg=CoxA=CoxLxW. Therefore Cg can be scaled by ßx1/ άx1/ ά= ß/ $ά^2$

4.Parasitic capacitance

Cx=Ax/d, where Ax is the area of the depletion around the drain or source. d is the depletion width .Ax is scaled down by 1/ά2 and d is scaled by 1/ά. Hence Cx is scaled by

1/ $ά^2$ /1/ ά = 1/ ά

5.Carrier density in the channel Qon

Qon=Co.Vgs

Co is scaled by ß and Vgs is scaled by 1/ ß,hence Qo is scaled by ßx1/ß =1.

Channel resistance Ro

Ron = L/Wx1/Qoxμ, μ is mobility of charge carriers . Ro is scaled by1/ά/1/ άx1=1

Gate delay Td

Td is proportional to Ro and Cg

Td is scaled by 1x ß/$ά^2$ = ß/$ά^2$

Maximum operating frequency fo fo=1/td,therefore it is scaled by 1/ ß/$ά^2$ = $ά^2$/ß **Saturation current**

Idss= CoμW(Vgs-Vt)/2L, Co scale by ß and voltages by 1/ ß, Idss is scaled by ß /$ß^2$= 1/ß **Current Density**

J=Idss/A hence J is scaled by 1/ß/1/ά2 = $ά^2$ /ß

UNIT - 5

SCALING OF MOS CIRCUITS: Scaling model and scaling factors- Limit due to current density.

1. What is scaling?
2. Why scaling?
3. Figure(s) of Merit (*FoM*) for scaling
4. International Technology Roadmap for Semiconductors (ITRS)
5. Scaling models
6. Scaling factors for device parameters
7. Implications of scaling on design
8. Limitations of scaling
9. Observations
10. Summary

Scaling of MOS Circuits

1.What is Scaling?

Proportional adjustment of the dimensions of an electronic device while maintaining the electrical properties of the device, results in a device either *larger* or *smaller* than the un-scaled device. Then *Which way do we scale the devices for VLSI? BIG and SLOW … or* ***SMALL*** *and* ***FAST****? What do we gain?*

2.Why Scaling?...

Scale the devices and wires down, Make the chips 'fatter' – functionality, intelligence, memory – and – faster, Make more chips per wafer – increased yield, Make the end user Happy by giving more for less and therefore, make MORE MONEY!!

3.FoM for Scaling

Impact of scaling is characterized in terms of several indicators:

- Minimum feature size
- Number of gates on one chip
- Power dissipation
- Maximum operational frequency
- Die size
- Production cost

Many of the FoMs can be improved by shrinking the dimensions of transistors and interconnections. Shrinking the separation between features – transistors and wires Adjusting doping levels and supply voltages.

3.1 Technology Scaling

Goals of scaling the dimensions by 30%:

Reduce gate delay by 30% (increase operating frequency by 43%)

Double transistor density

Reduce energy per transition by 65% (50% power savings @ 43% increase in frequency)

Die size used to increase by 14% per generation

Technology generation spans 2-3 years

Figure1 to Figure 5 illustrates the technology scaling in terms of minimum feature size, transistor count, prapogation delay, power dissipation and density and technology generations.

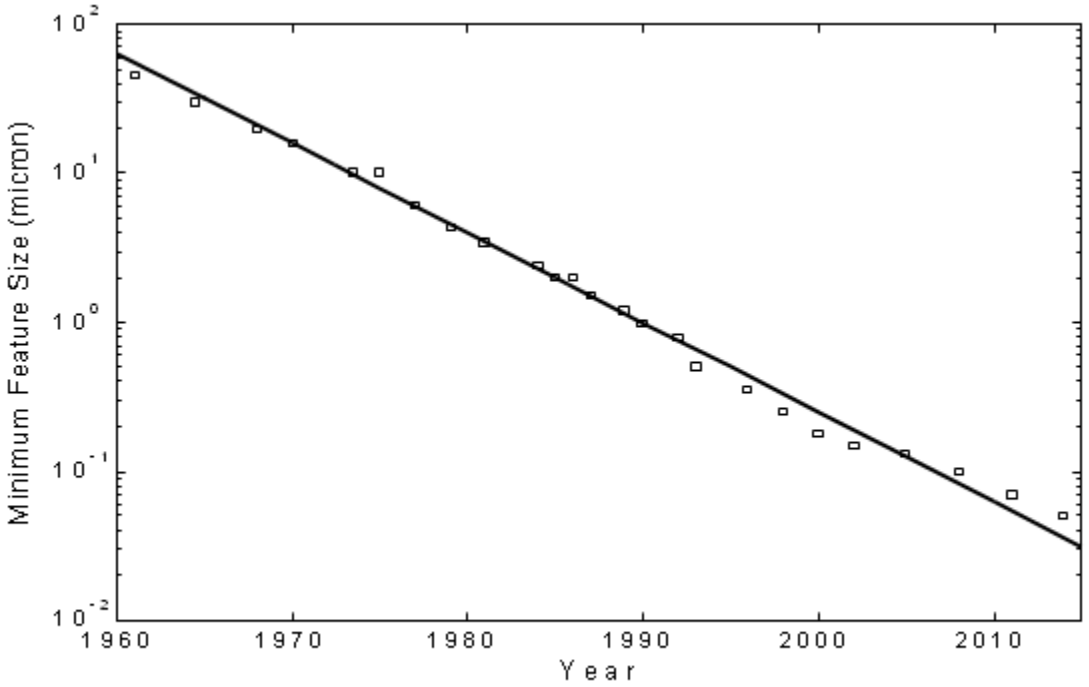


Figure-1:Technology Scaling (1)

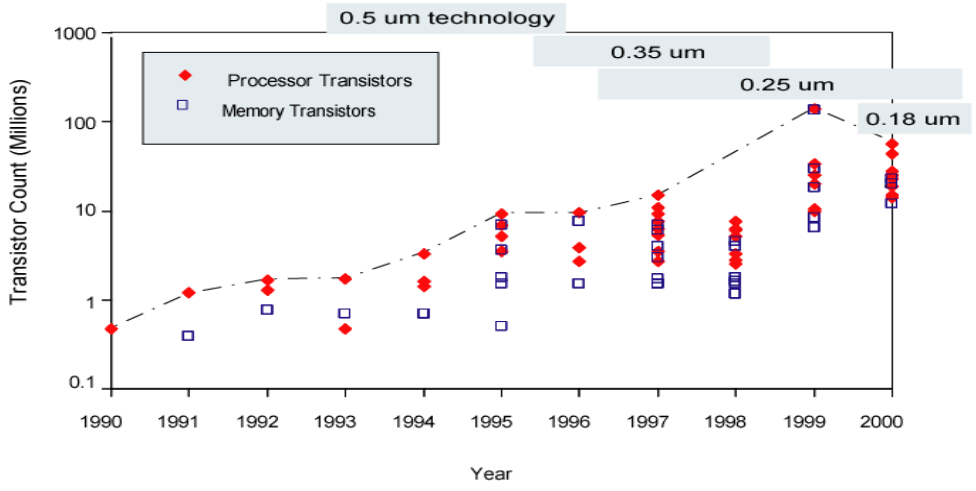


Figure-2:Technology Scaling (2)

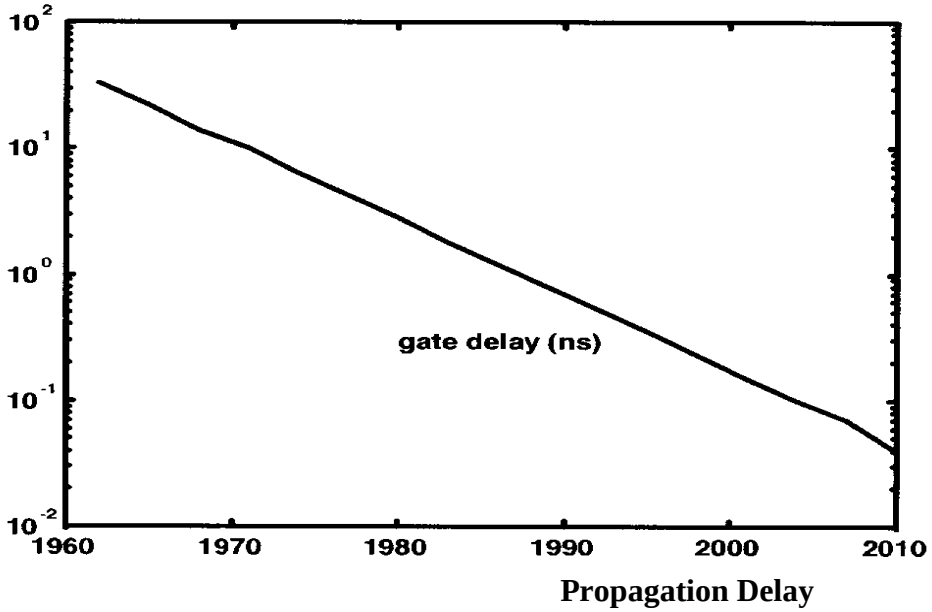
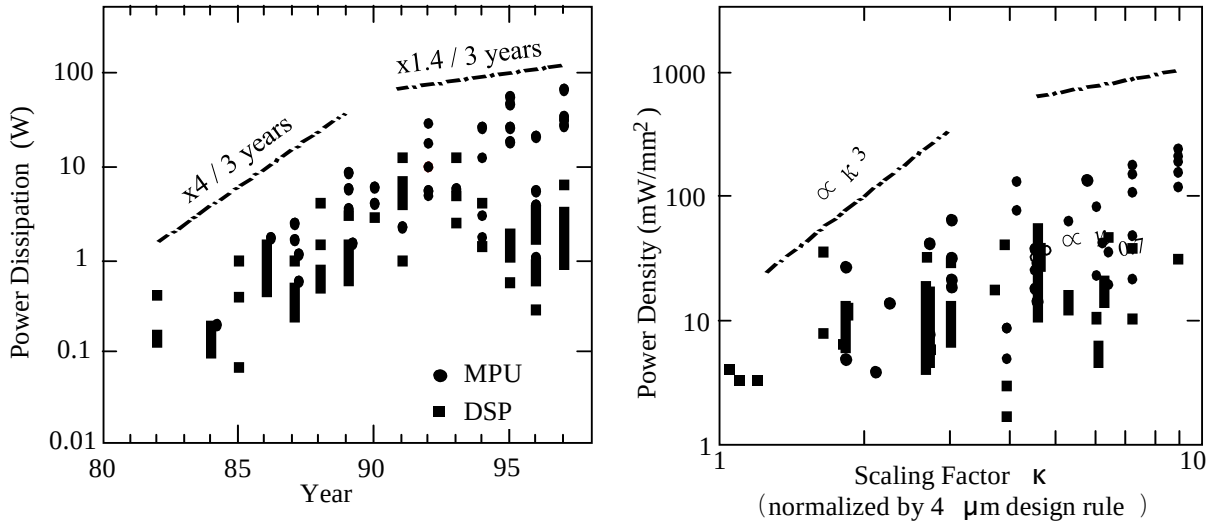


Figure-3:Technology Scaling (3)



(a) Power dissipation vs. year. (b) Power density vs. scaling factor.

Figure-4:Technology Scaling (4)

Technology Generations

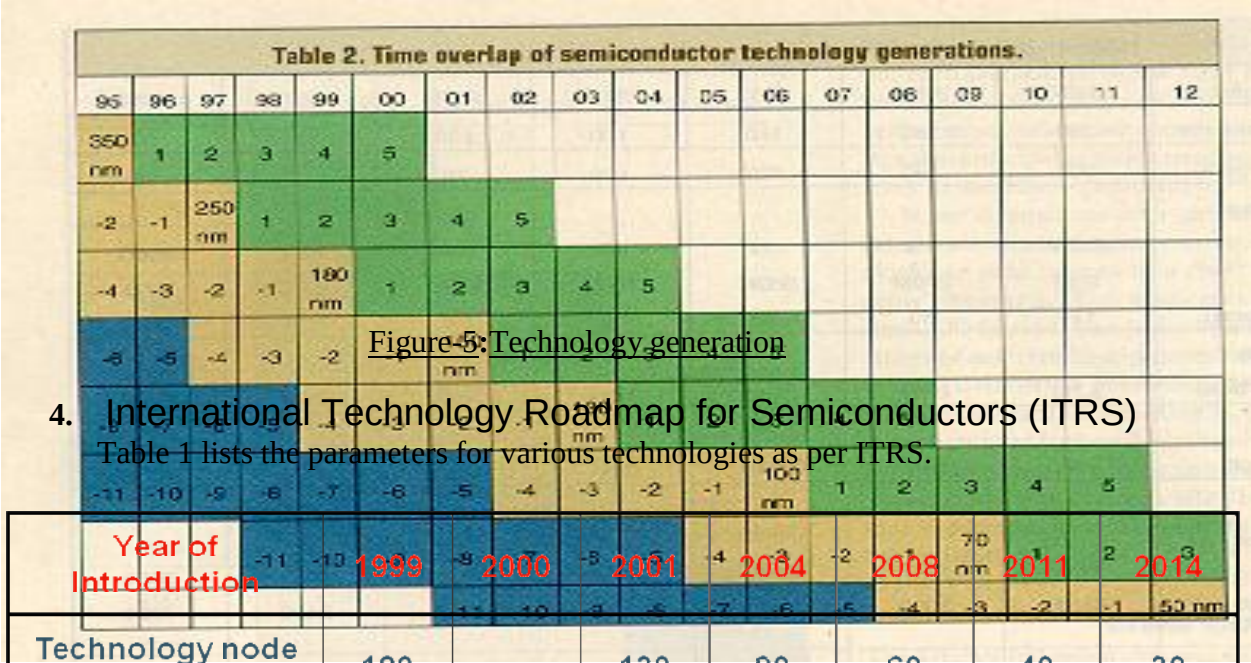


Figure-5: Technology generation

4. International Technology Roadmap for Semiconductors (ITRS)

Table 1 lists the parameters for various technologies as per ITRS.

Year of Introduction	1999	2000	2001	2004	2008	2011	2014
Technology node [nm]	180		130	90	60	40	30
Supply [V]	1.5-1.8	1.5-1.8	1.2-1.5	0.9-1.2	0.6-0.9	0.5-0.6	0.3-0.6
Wiring levels	6-7	6-7	7	8	9	9-10	10
Max frequency [GHz], Local-Global	1.2	1.6-1.4	2.1-1.6	3.5-2	7.1-2.5	11-3	14.9 -3.6
Max µP power [W]	90	106	130	160	171	177	186
Bat. power [W]	1.4	1.7	2.0	2.4	2.1	2.3	2.5

Table 1: ITRS

2010/45nm. 2016/23nm.
Node years: 2007/65nm.
2013/33nm.

5.Scaling Models

Full Scaling (Constant Electrical Field)

Ideal model – dimensions and voltage scale together by the same scale factor

Fixed Voltage Scaling

Most common model until recently – only the dimensions scale, voltages remain constant

General Scaling

Most realistic for today's situation – voltages and dimensions scale with different factors

6.Scaling Factors for Device Parameters

Device scaling modeled in terms of generic scaling factors:
1/α and 1/β

- 1/β: scaling factor for supply voltage V_{DD} and gate oxide thickness D
- 1/α: linear dimensions both horizontal and vertical dimensions

Why is the scaling factor for gate oxide thickness different from other linear horizontal and vertical dimensions? Consider the cross section of the device as in Figure 6,various parameters derived are as follows.

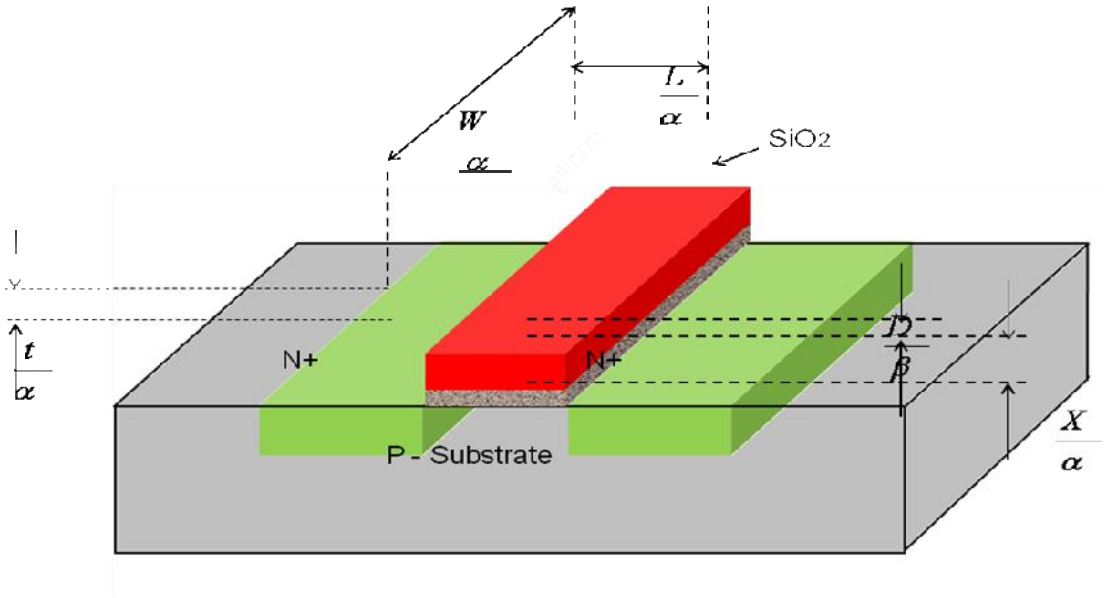


Figure-6:Technology generation

Gate area A_g

$$A_g = L*W$$

Where L: Channel length and W: Channel width and both are scaled by 1/α

Thus A_g is scaled up by $1/\alpha^2$

- Gate capacitance per unit area C_o or C_{ox}

$C_{ox} = \varepsilon_{ox}/D$

Where ε_{ox} is permittivity of gate oxide(thin-ox)= $\varepsilon_{ins}\varepsilon_o$ and D is the gate oxide thickness scaled by 1/β Thus C_{ox} is scaled up by

1 = β

1

β

- Gate capacitance C_g $C_g = C_o * L * W$

Thus C_g is scaled up 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 and scaled by $1/\alpha$

A_x is the area of the depletion region around source or drain, scaled by $(1/\alpha^2)$.
Thus C_x is scaled up by $\{1/(1/\alpha)\} * (1/\alpha^2) = 1/\alpha$

- Carrier density in channel Q_{on}

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

where Q_{on} is the average charge per unit area 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

$$R_{on} = \frac{L}{W} * \frac{1}{Q_{on} * \mu}$$

- Channel Resistance ~~R_{on}~~

Where μ = channel carrier mobility and assumed constant

Thus R_{on} is scaled by 1.

- Gate delay T_d

T_d is proportional to $R_{on}*C_g$

T_d is scaled by $\frac{1}{\alpha^2} * \beta = \frac{\beta}{\alpha^2}$

- Maximum operating frequency f_o

$$f_o = \frac{W}{L} * \frac{\mu C_o V_{DD}}{C_g}$$

f_o is inversely proportional to delay T_d and is scaled by

$$\frac{1}{\frac{1}{\beta/\alpha^2}} = \frac{\alpha^2}{\beta}$$

- Saturation current I_{dss}

$$I_{dss} = \frac{C_o \mu}{2} * \frac{W}{L} * (V_{gs} - V_t)^2$$

Both V_{gs} and V_t are scaled by $(1/\beta)$. Therefore, I_{dss} is scaled by $\frac{1}{\beta/\alpha^2} = \frac{\alpha^2}{\beta}$

- Power dissipation per gate P_g

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

P_g comprises of two components: static component P_{gs} and dynamic component P_{gd}:

Where, the static power component is given by: $P_{gs} = \frac{V_{DD}^2}{R_{on}}$

And the dynamic component by: $P_{gd} = E_g f_o$

Since V_{DD} scales by $(1/\beta)$ and R_{on} scales by 1, P_{gs} scales by $(1/\beta^2)$.

Since Eg scales by $(1/\alpha^2\ \beta)$ and f_o by $(\alpha_2\ /\beta)$, P_{gd} also scales by $(1/\beta^2)$. Therefore, P_g scales by $(1/\beta^2)$.

- Power dissipation per unit area P_a

6.1 Scaling Factors …Summary

Various device parameters for different scaling models are listed in Table 2 below.

Table 2: Device parameters for scaling models

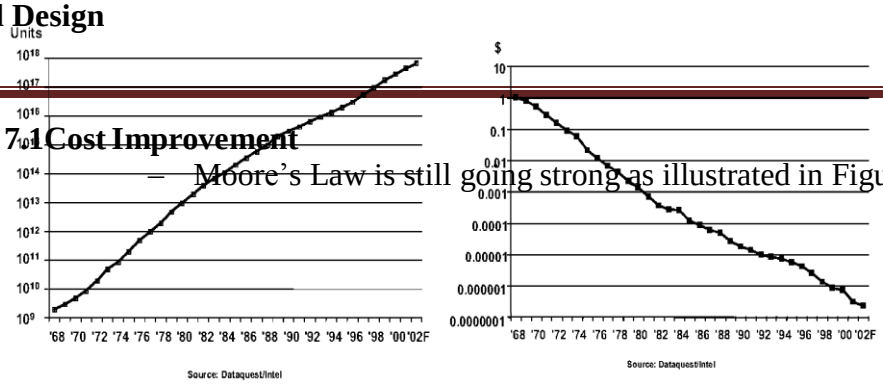
NOTE: for Constant E: β=α; for Constant V: β=1

Parameters	Description	General (Combined V and Dimension)	Constant E	Constant V
V_{DD}	Supply voltage	1/β	1/α	1
L	Channel length	1/α	1/α	1/α
W	Channel width	1/α	1/α	1/α
D	Gate oxide thickness	1/β	1/α	1
A_g	Gate area	$1/\alpha^2$	$1/\alpha^2$	$1/\alpha^2$
C_o (or C_{ox})	Gate capacitance per unit area	β	α	1
C_g	Gate capacitance	β/α^2	1/α	$1/\alpha^2$
C_x	Parsitic capacitance	1/α	1/α	1/α
Q_{on}	Carrier density	1	1	1
R_{on}	Channel resistance	1	1	1
I_{dss}	Saturation current	1/β	1/α	1

Parameters	Description	General (Combined V and Dimension)	Constant E	Constant V
A_c	Conductor cross section area	$1/\alpha^2$	$1/\alpha^2$	$1/\alpha^2$
J	Current density	α^2 / β	α	α^2
V_g	Logic 1 level	$1 / \beta$	$1 / \alpha$	1
E_g	Switching energy	$1 / \alpha^2 \beta$	$1 / \alpha^3$	$1/\alpha^2$
P_g	Power dissipation per gate	$1 / \beta^2$	$1/\alpha^2$	1
N	Gates per unit area	α^2	α^2	α^2
P_a	Power dissipation per unit area	α^2 / β^2	1	α^2
T_d	Gate delay	β / α^2	$1 / \alpha$	$1/\alpha^2$
f_o	Max. operating frequency	α^2 / β	α	α^2
P_T	Power speed product	$1 / \alpha^2 \beta$	$1 / \alpha^3$	$1/\alpha^2$

7.Implications of Scaling

Improved Performance
Improved Cost
Interconnect Woes
Power Woes
Productivity Challenges
Physical Limits



7.1Cost Improvement

– Moore's Law is still going strong as illustrated in Figure 7.

Figure-7:Technology generation

7.2:Interconnect Woes

- Scaled transistors are steadily improving in delay, but scaled wires are holding constant or getting worse.
- SIA made a gloomy forecast in 1997
  - Delay would reach minimum at 250 – 180 nm, then get worse because of wires
- But…
- For short wires, such as those inside a logic gate, the wire RC delay is negligible.
- However, the long wires present a considerable challenge.

- Scaled transistors are steadily improving in delay, but scaled wires are holding constant or getting worse.
- SIA made a gloomy forecast in 1997
  - Delay would reach minimum at 250 – 180 nm, then get worse because of wires
- But…
- For short wires, such as those inside a logic gate, the wire RC delay is negligible.
- However, the long wires present a considerable challenge.

Figure 8 illustrates delay Vs. generation in nm for different materials.

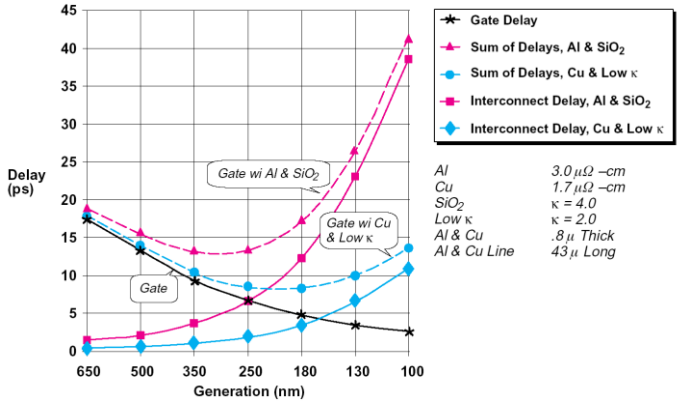


Figure-8:Technology generation

7.3 Reachable Radius

- We can't send a signal across a large fast chip in one cycle anymore
- But the microarchitect can plan around this as shown in Figure 9.
  - Just as off-chip memory latencies were tolerated

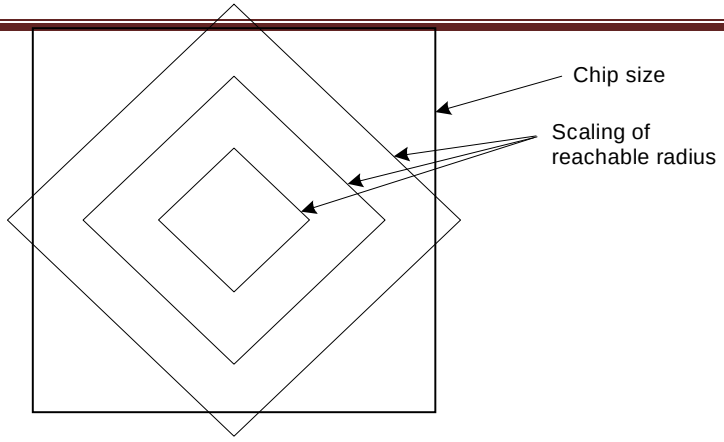


Figure-9**:**Technology generation

7.4 Dynamic Power

- Intel VP Patrick Gelsinger (ISSCC 2001)
  - If scaling continues at present pace, by 2005, high speed processors would have power density of nuclear reactor, by 2010, a rocket nozzle, and by 2015, surface of sun.
  - "Business as usual will not work in the future."
- Attention to power is increasing(Figure 10)

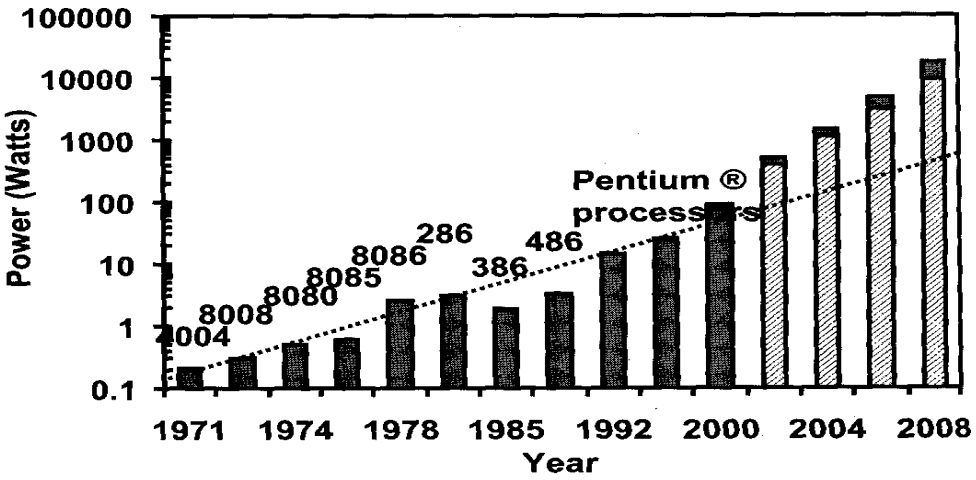
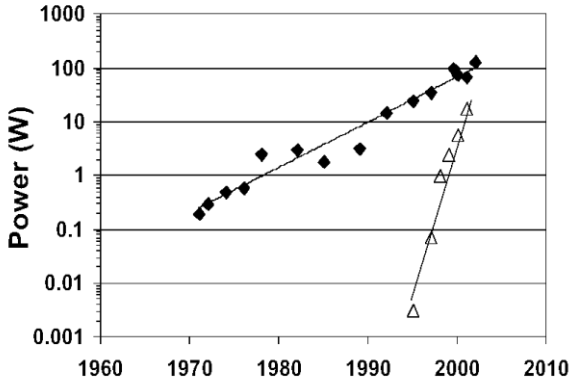


Figure-10:Technology generation

7.5 Static Power

- V_{DD} decreases
  - Save dynamic power
  - Protect thin gate oxides and short channels
  - No point in high value because of velocity saturation.
- V_t must decrease to maintain device performance
- But this causes exponential increase in OFF leakage
  A Major future challenge(Figure 11)



Moore(03) Figure-11:Technology generation

7.6 Productivity

- Transistor count is increasing faster than designer productivity (gates / week)
  - Bigger design teams
    - Up to 500 for a high-end microprocessor
  - More expensive design cost
  - Pressure to raise productivity
    - Rely on synthesis, IP blocks
  - Need for good engineering managers

7.7 Physical Limits

- Will Moore's Law run out of steam?

Can't build transistors smaller than an atom…

- Many reasons have been predicted for end of scaling
  - Dynamic power
  - Sub-threshold leakage, tunneling
  - Short channel effects
  - Fabrication costs
  - Electro-migration
  - Interconnect delay
- Rumors of demise have been exaggerated

8. Limitations of Scaling

Effects, as a result of scaling down- which eventually become severe enough to prevent further miniaturization.

- Substrate doping
- Depletion width
- Limits of miniaturization
- Limits of interconnect and contact resistance
- Limits due to sub threshold currents
- Limits on logic levels and supply voltage due to noise
- Limits due to current density

8.1 Substrate doping

- Substrate doping
- Built-in(junction) potential V_B depends on substrate doping level – can be neglected as long as V_B is small compared to $V_{DD.}$
- As length of a MOS transistor is reduced, the depletion region width –scaled down to prevent source and drain depletion region from meeting.
- the depletion region width d for the junctions is $d =$ _____
- ε_{si} relative permittivity of silicon

8.2 Depletion width

- N_B is increased to reduce d , but this increases threshold voltage V_t -against trends for scaling down.
- Maximum value of N_B ($1.3*10^{19}$ cm^{-3} , at higher values, maximum electric field applied to gate is insufficient and no channel is formed.
- N_B maintained at satisfactory level in the channel region to reduce the above problem.

- E_{max} maximum electric field induced in the junction.
- Electric field across the depletion region is increased by

  1/ $\sqrt{\frac{\ln\alpha}{\alpha}}$

- Reach a critical level E_{crit} with increasing N_B $d = \sqrt{\frac{2\xi_{si}}{q N_B}\left(E_{crit}.d\right)}$

Where $d = \frac{\xi_{si}\,\xi_0}{q\,N_B}(E_{crit})$

Figure 12 , Figure 13 and Figure 14 shows the relation between substrate concentration Vs depletion width , Electric field and transit time.
Figure 15 demonstrates the interconnect length Vs. propagation delay and Figure 16 oxide thickness Vs. thermal noise.

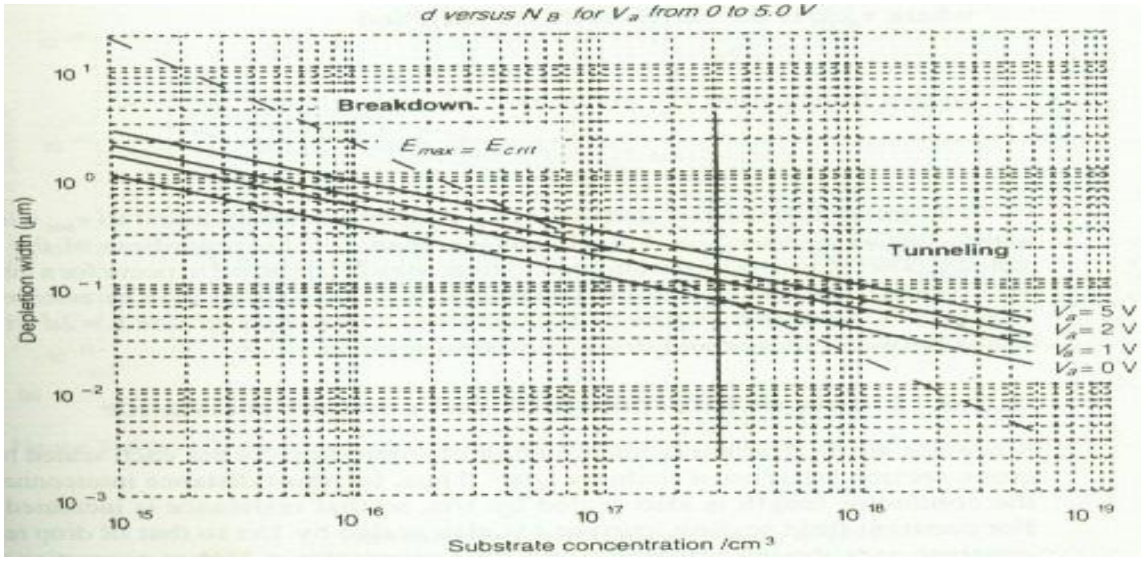


Figure-12:Technology generation

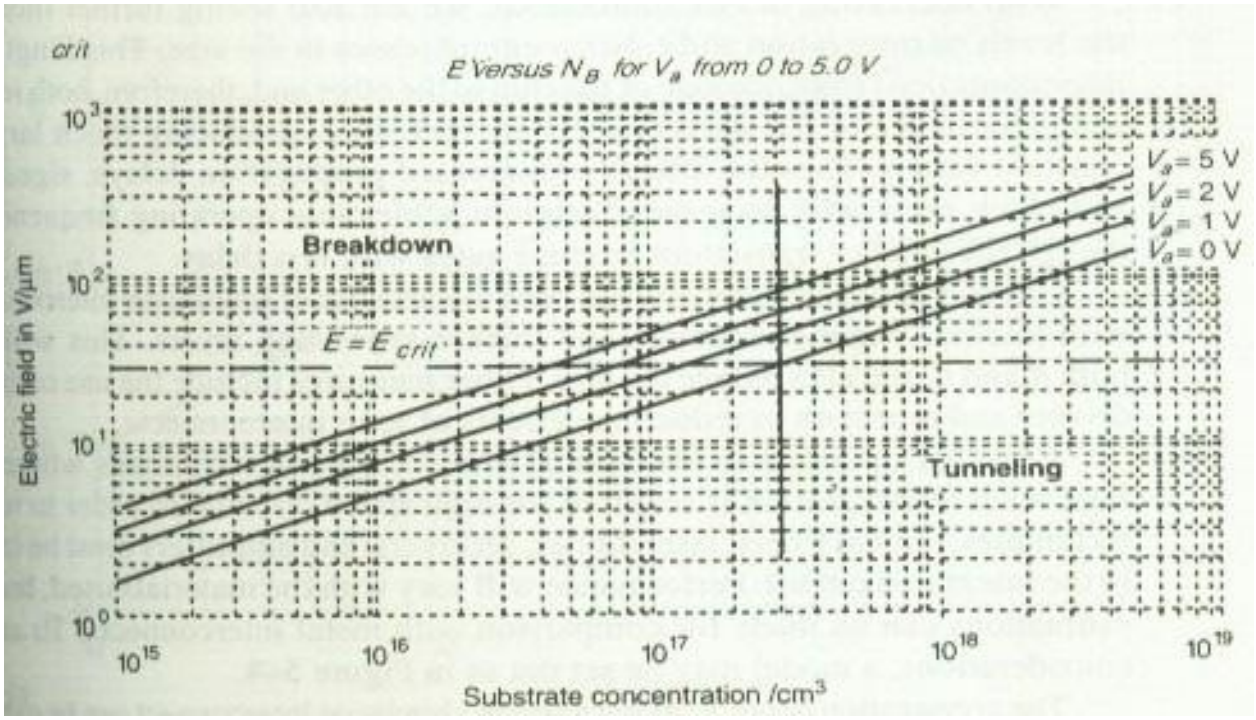


Figure-13:Technology generation

8.3 Limits of miniaturization

- minimum size of transistor; process tech and physics of the device
- Reduction of geometry; alignment accuracy and resolution

- Size of transistor measured in terms of channel length L

  L=2d (to prevent push through)
- L determined by N_B and Vdd
- Minimum transit time for an electron to travel from source to drain is

$$v_{drift} = \mu E$$

$$t = \frac{L}{} = \frac{2d}{}$$

V_{drift} μE

- ₛmaximum carrier drift velocity is approx. Vsat,regardless of supply voltage

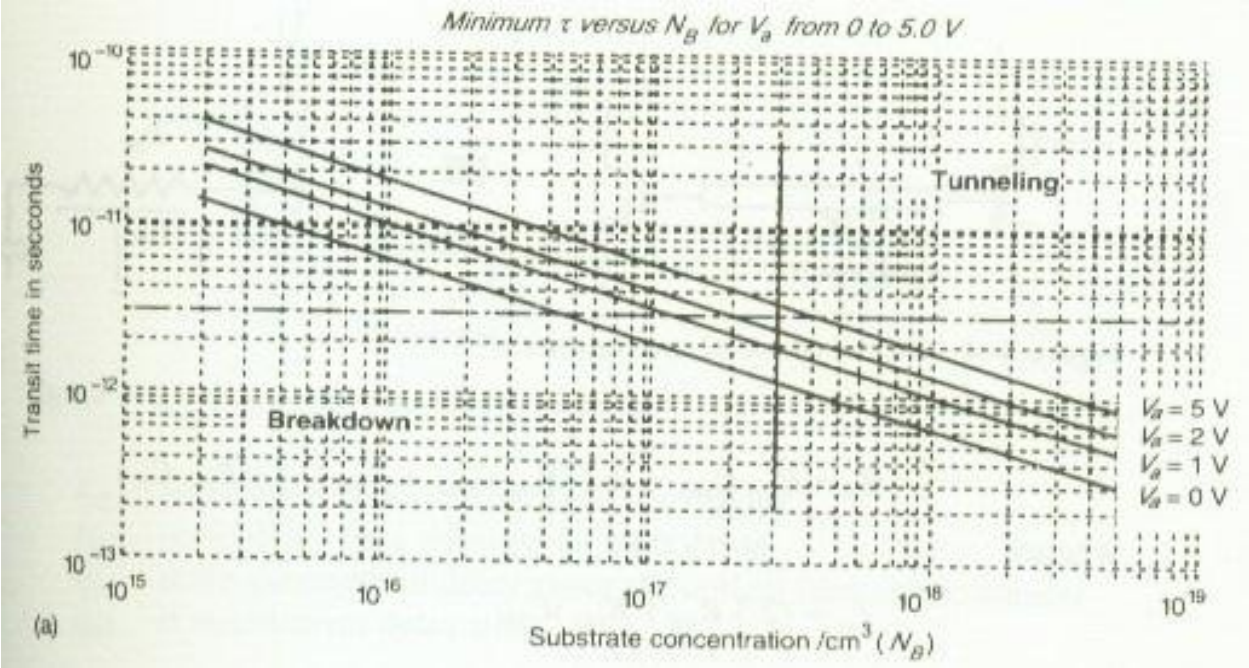


Figure-14:Technology generation

8.4 Limits of interconnect and contact resistance

- Short distance interconnect- conductor length is scaled by 1/α and resistance is increased by α
- For constant field scaling, I is scaled by 1/ α so that IR drop remains constant as a result of scaling.-driving capability/noise margin.

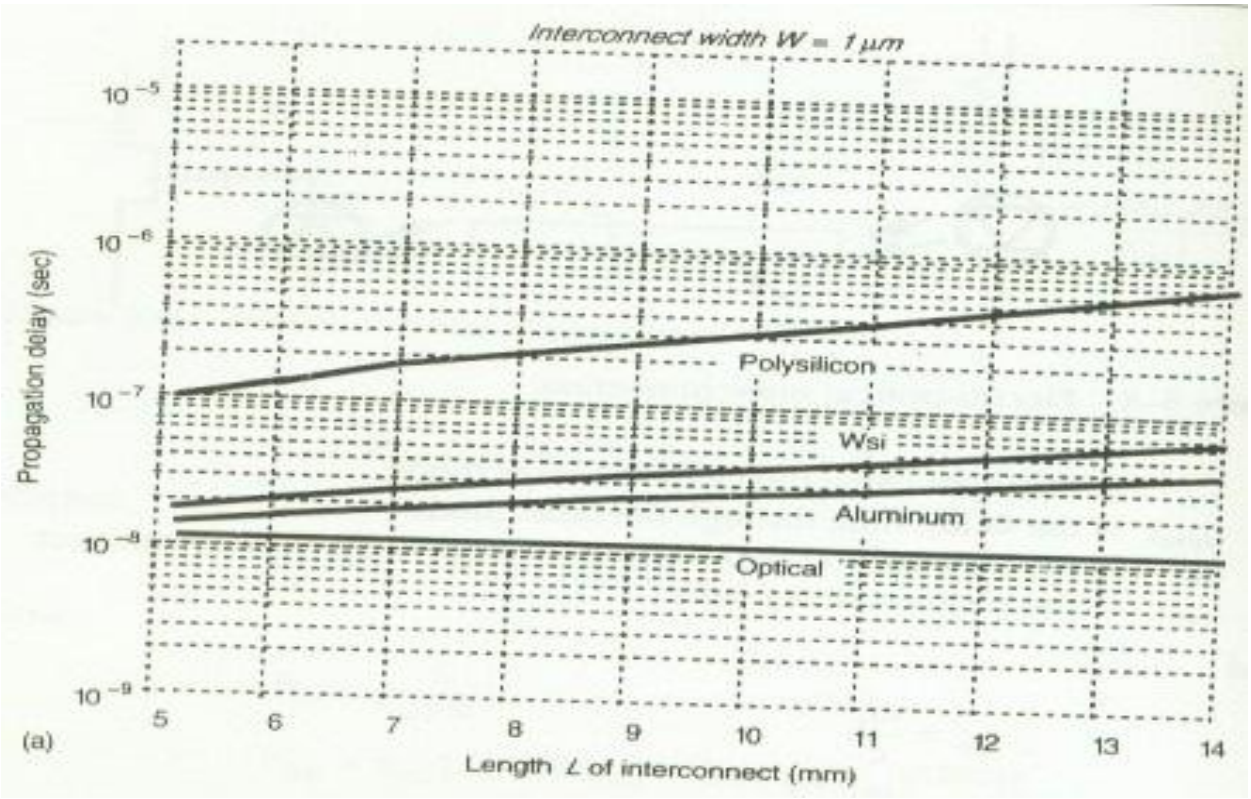


Figure-15:Technology generation

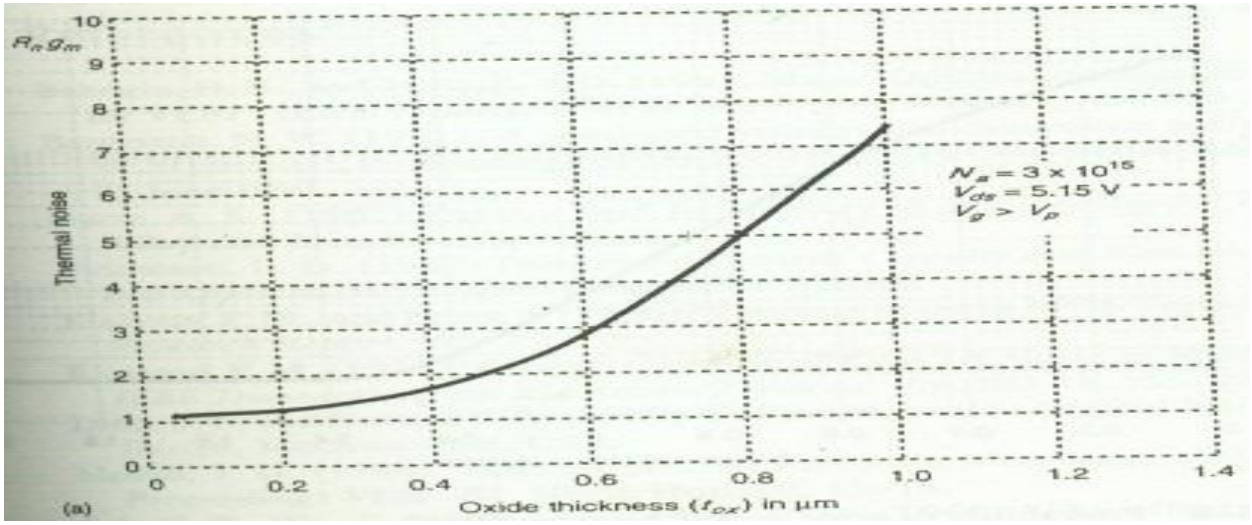
8.5 Limits due to subthreshold currents

- Major concern in scaling devices.
- I_{sub} is directly praportinal exp (Vgs – Vt) q/KT
- As voltages are scaled down, ratio of Vgs-Vt to KT will reduce-so that threshold current increases.
- Therefore scaling Vgs and Vt together with Vdd .
- Maximum electric field across a depletion region is

$$E_{\max} = 2\{V_a + V_b\}/d$$

8.6 Limits on supply voltage due to noise

Decreased inter-feature spacing and greater switching speed –result in noise problems



9. Observations – Device scaling

- Gate capacitance per micron is nearly independent of process
- But ON resistance * micron improves with process
- Gates get faster with scaling (good)
- Dynamic power goes down with scaling (good)
- Current density goes up with scaling (bad)
- Velocity saturation makes lateral scaling unsustainable

UNIT - 6

SUBSYSTEM DESIGN AND LAYOUT: Some architecture issues- other systems considerations. Examples of structural design, clocked sequential circuits

8 Hours

CMOS SUBSYSTEM DESIGN
CONTENTS

1. System
2. VLSI design flow
3. Structured design approach
4. Architectural issues
5. MOSFET as switch for logic functionality
6. Circuit Families
   Restoring Logic: CMOS and its variants - NMOS and Bi CMOS Other circuit variants
   NMOS gates with depletion (zero -threshold) pull up
   Bi-CMOS gates

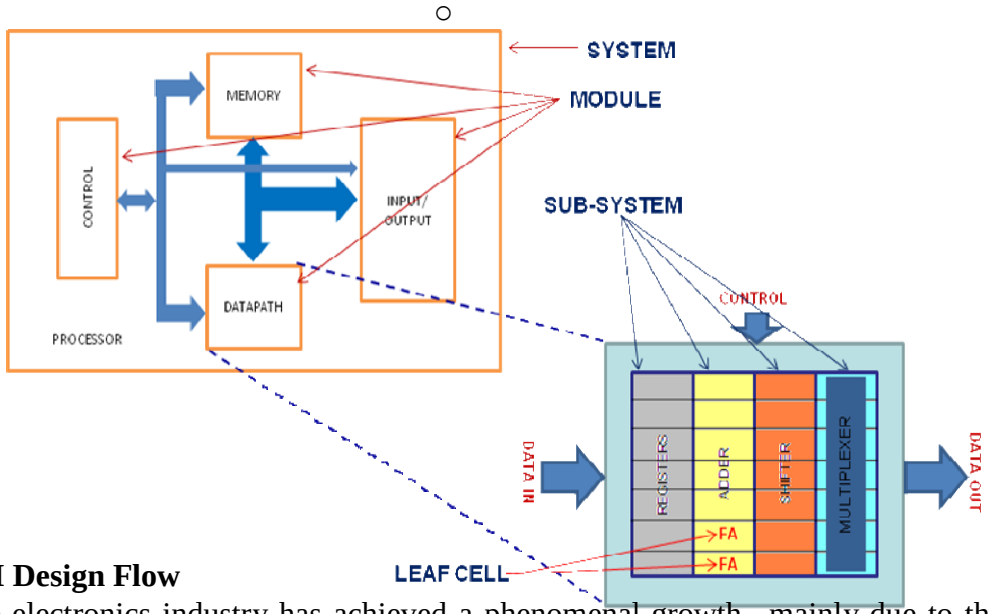
7. Switch logic: Pass Transistor and Transmission gate (TG)
8. Examples of Structured Design
   MUX
   DMUX
   D Latch and Flop
   A general logic function block

1.What is a *System*?

A *system* is a set of interacting or interdependent entities forming and integrate whole. Common characteristics of a system are

- Systems have *structure* - defined by parts and their composition
- Systems have *behavior* – involves inputs, processing and outputs (of material, information or energy)
- Systems have *interconnectivity* the various parts of the system functional as well as structural relationships between each other

1.1Decomposition of a *System*: A Processor



5.VLSI Design Flow

- The electronics industry has achieved a phenomenal growth –mainly due to the rapid advances in integration technologies, large scale systems design-in short due to VLSI.
- Number applications of integrated circuits in high-performance computing, telecommunications, and consumer electronics has been rising steadily.
- Current leading-edge technology trend –expected to continue with very important implications on VLSI and systems design.
- The design process, at various levels, is evolutionary in nature.
- Y-Chart (first introduced by D. Gajski) as shown in Figure1 illustrates the design flow for mast logic chips, using design activities.
- Three different axes (domains) which resemble the letter Y.
- Three major domains, namely
  Behavioral domain
  Structural domain

Geometrical domain

- Design flow starts from the algorithm that describes the behavior of target chip.

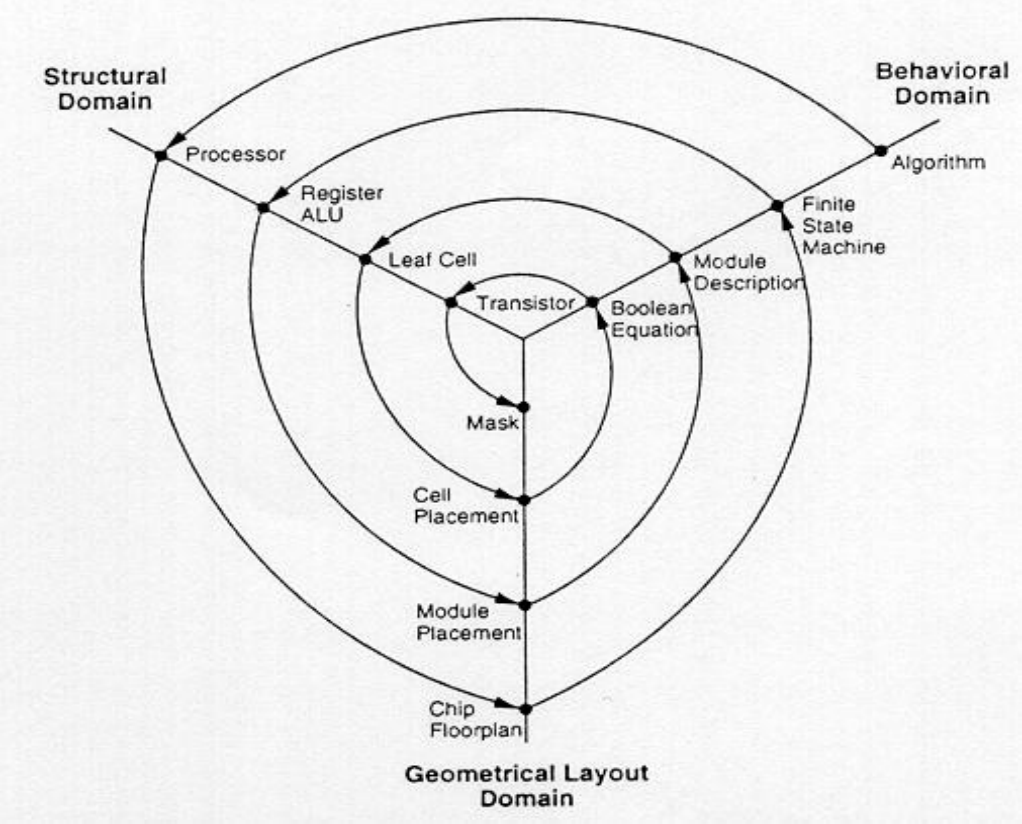


Figure 1. Typical VLSI design flow in three domains(Y-chart)

VLSI design flow, taking in to account the various representations, or abstractions of design are Behavioural,logic,circuit and mask layout.

Verification of design plays very important role in every step during process.

Two approaches for design flow as shown in Figure 2 are

Top-down

Bottom-up

Top-down design flow- excellent design process control

In reality, both top-down and bottom-up approaches have to be combined.

Figure 3 explains the typical full custom design flow.

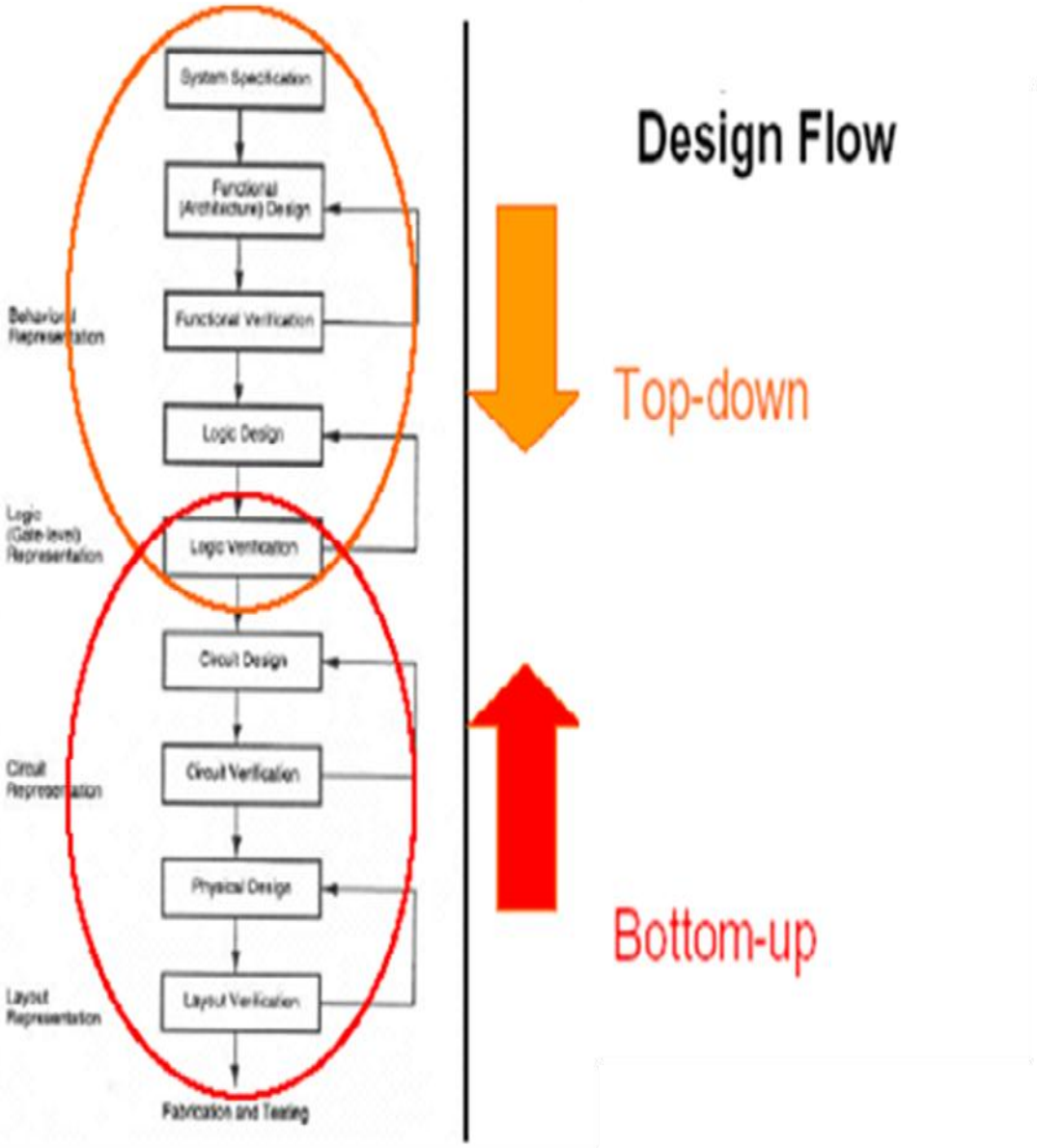


Figure 2. Typical VLSI design flow

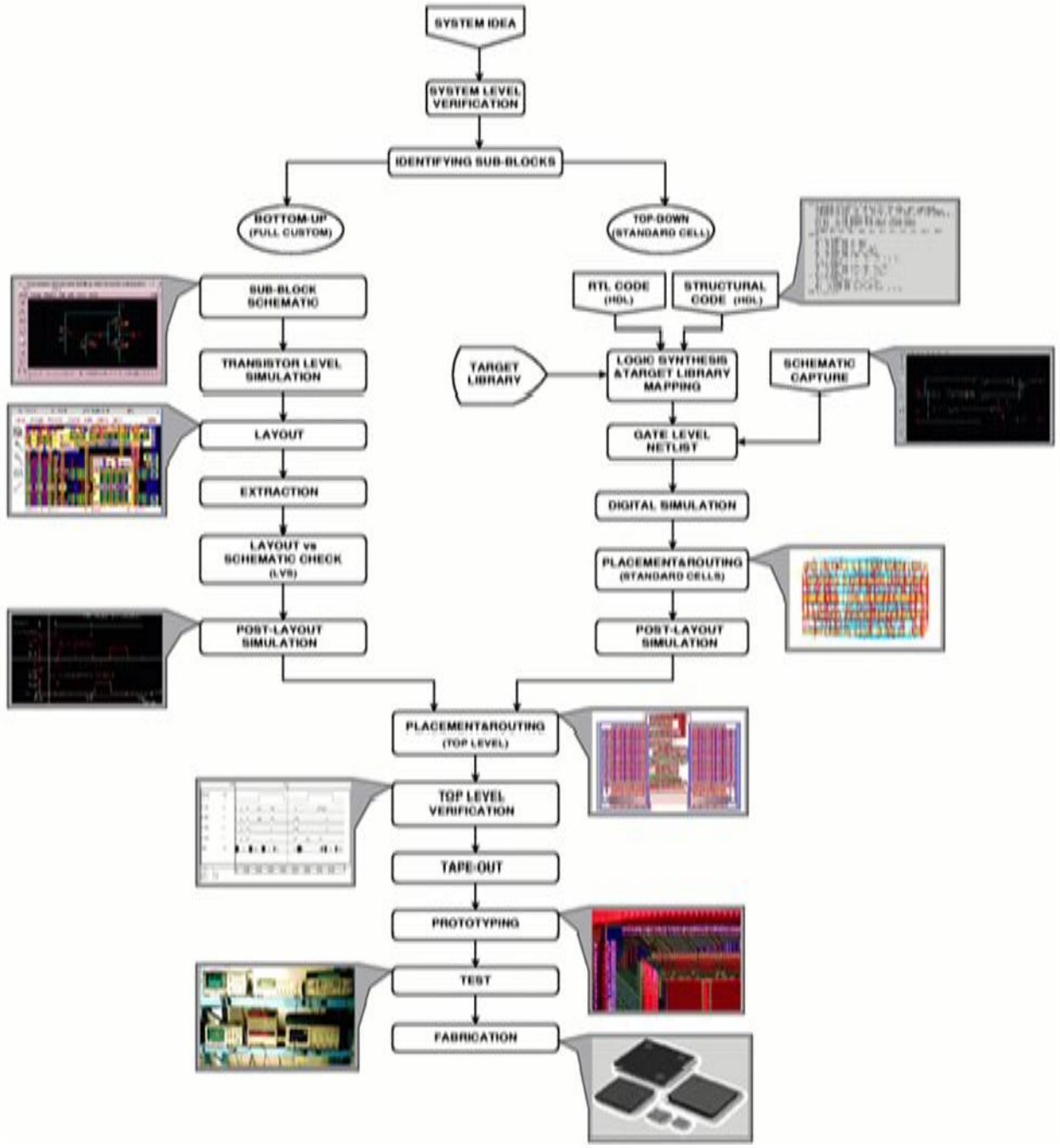


Figure 3. Typical ASIC/Custom design flow

3 Structured Design Approach

- Design methodologies and structured approaches developed with complex hardware and software.
- Regardless of the actual size of the project, basic principles of structured design- improve the prospects of success.
- Classical techniques for reducing the complexity of IC design are:
  Hierarchy
  Regularity
  Modularity
  Locality

Hierarchy: "Divide and conquer" technique involves dividing a module into sub-modules and then repeating this operation on the sub-modules until the complexity of the smaller parts becomes manageable.

Regularity: The hierarchical decomposition of a large system should result in not only **simple**, but also **similar** blocks, as much as possible. Regularity usually reduces the number of different modules that need to be designed and verified, at all levels of abstraction.

Modularity: The various functional blocks which make up the larger system must have **well-defined functions** and **interfaces**.

Locality: Internal details remain at the local level. The concept of locality also ensures that connections are mostly between neighboring modules, **avoiding long-distance connections** as much as possible.

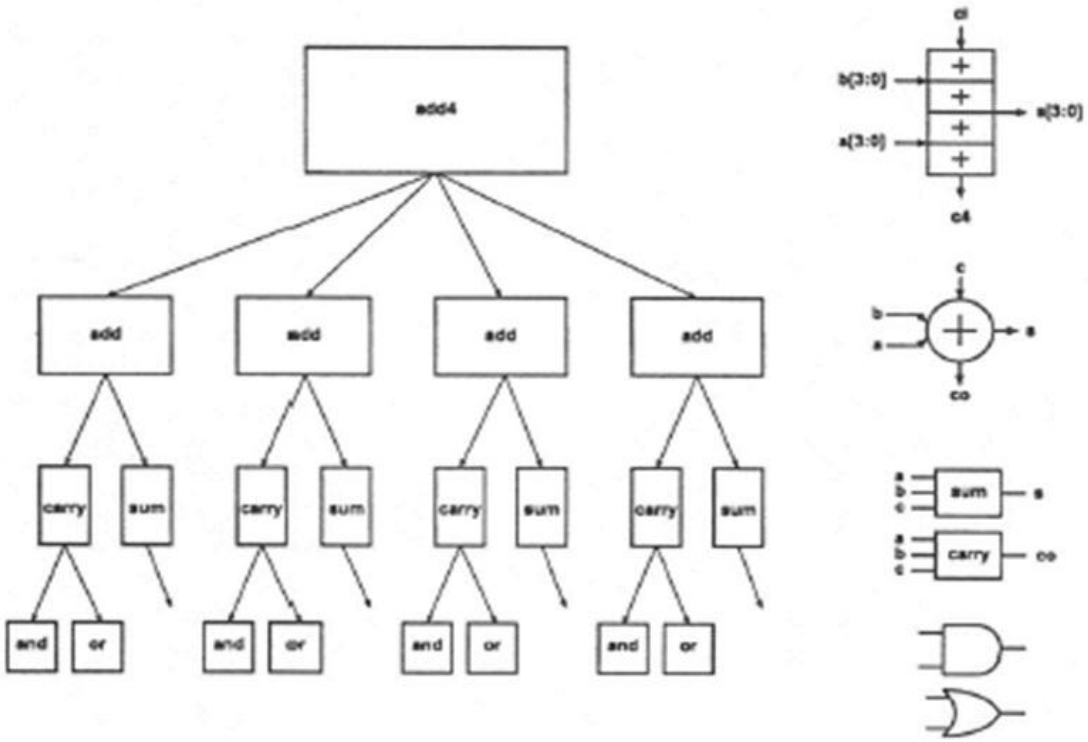


Figure 4-Structured Design Approach –Hierarchy

Regularity

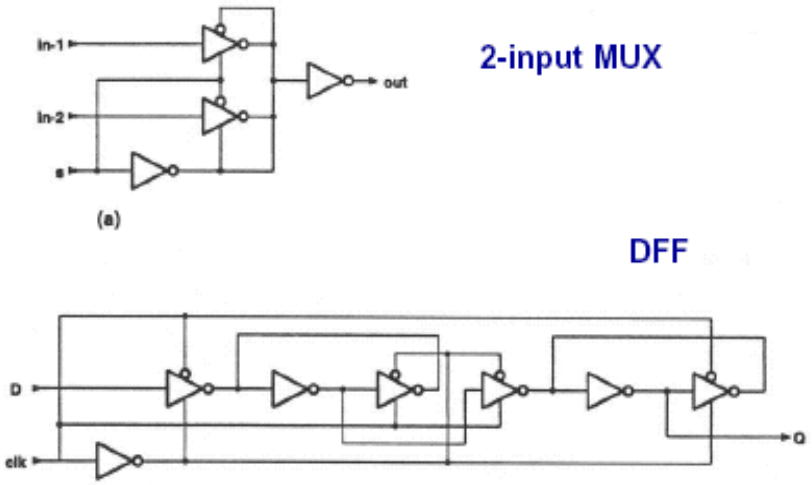


Figure5-.Structured Design Approach –Regularity

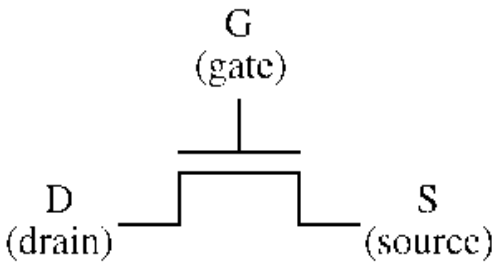
- Design of array structures consisting of identical cells.-such as parallel multiplication array.
- Exist at all levels of abstraction:
  transistor level-uniformly sized.
  logic level- identical gate structures
- 2:1 MUX, D-F/F- inverters and tri state buffers
- Library-well defined and well-characterized basic building block.
- Modularity: enables parallelization and allows plug-and-play
- Locality: Internals of each module unimportant to exterior modules and internal details remain at local level.

Figure 4 and Figure 5 illustrates these design approaches with an example.

4 Architectural issues

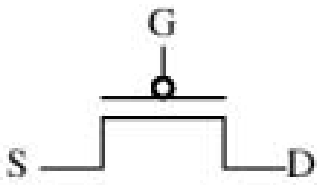
- Design time increases exponentially with increased complexity
- Define the requirements
- Partition the overall architecture into subsystems.
- Consider the communication paths
- Draw the floor plan
- Aim for regularity and modularity
- convert each cell into layout
- Carry out DRC check and simulate the performance

5. MOSFET as a Switch



*n*MOS transistor:
Closed (conducting) when
Gate = 1 (Vdd, 5V)

Open (non-conducting) when
Gate = 0 (ground, 0V)



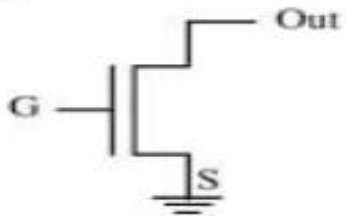
*p*MOS transistor:
Closed (conducting) when
Gate = 0 (ground, 0V)

Open (non-conducting) when
Gate = 1 (Vdd, 5V)

Note: The MOS transistor is a symmetric device. This means that the drain and source terminals are interchangeable. For a conducting *n*MOS transistor, $V_{DS} > 0V$; for the *p*MOS transistor, $V_{DS} < 0V$ (or $V_{SD} > 0V$).

- We can view MOS transistors as electrically controlled switches
- Voltage at gate controls path from source to drain

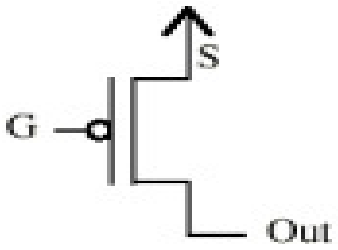
For *n*MOS switch, source is typically tied to ground and is used to *pull-down* signals:



when Gate = 1, Out = 0, (0V)

when Gate = 0, Out = Z (high impedance)

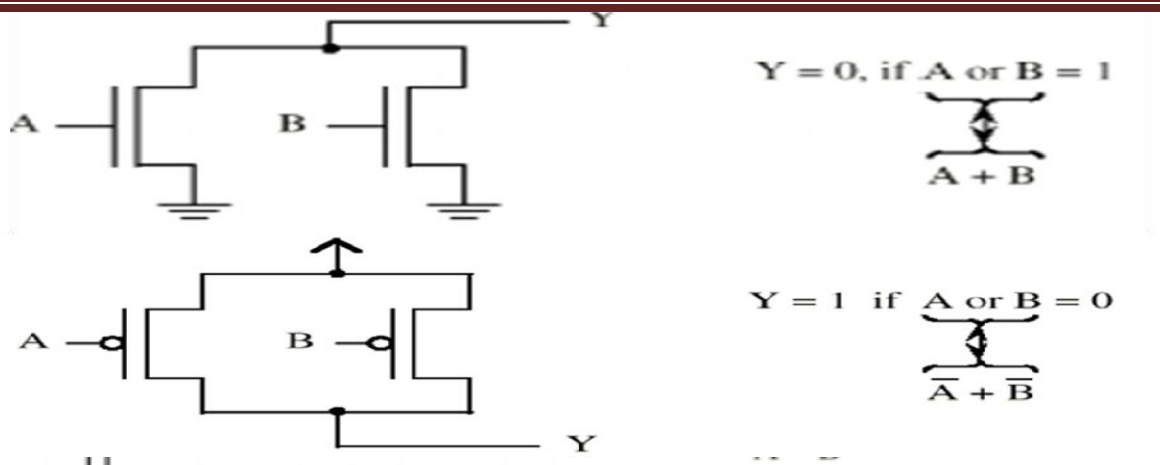
For *p*MOS switch, source is typically tied to Vdd, used to *pull* signals *up*:



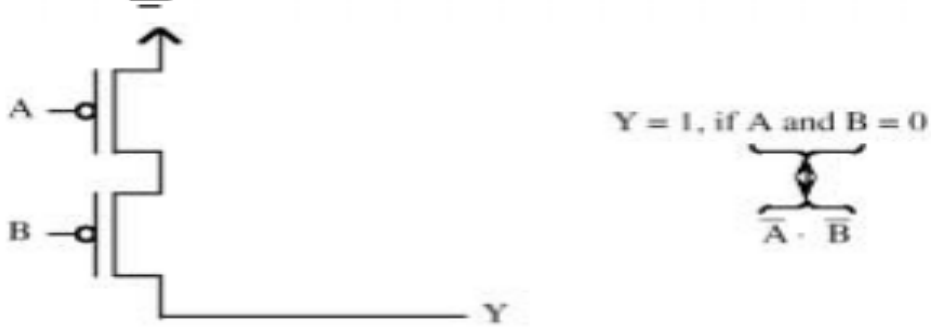
when Gate = 0, Out = 1 (Vdd)

when Gate = 1, Out = Z (high impedance)

5.1 Parallel connection of Switches..



5.2 Series connection of Switches..



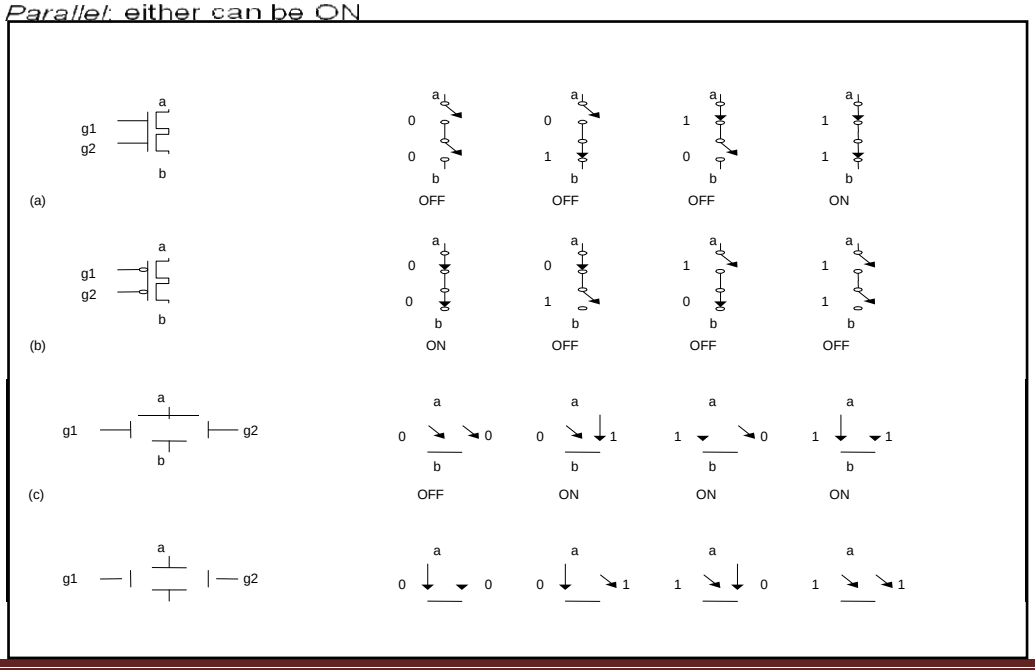
5.3 Series and parallel connection of Switches..

nMOS: 1 = ON

pMOS: 0 = ON

Series: both must be ON

Parallel: either can be ON

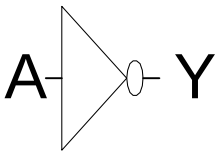


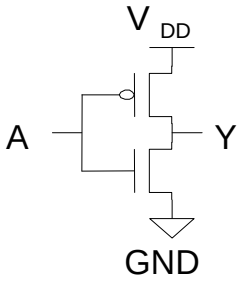
	b	b	b	b	b
(d)		ON	ON	ON	OFF

6. Circuit Families : Restoring logic

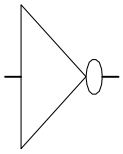
CMOS INVERTER

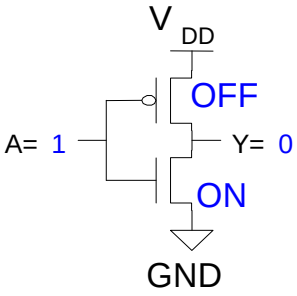
A	Y
0	
1	



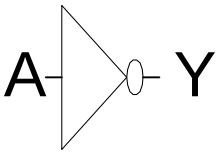


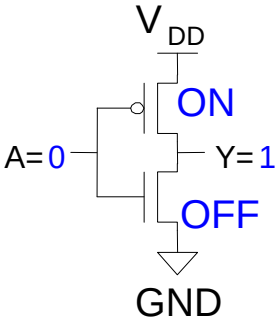
A	Y
0	
1	**0**





A	Y
0	**1**
1	0





6.1 NAND gate Design..

NAND Gate Design

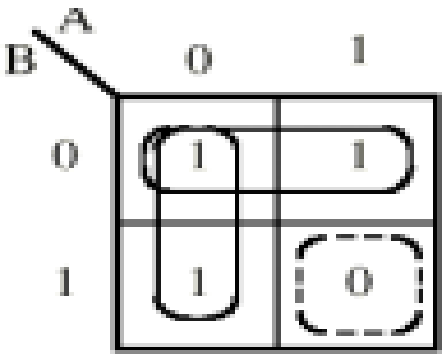
p-type transistor tree will provide "1" values of logic function

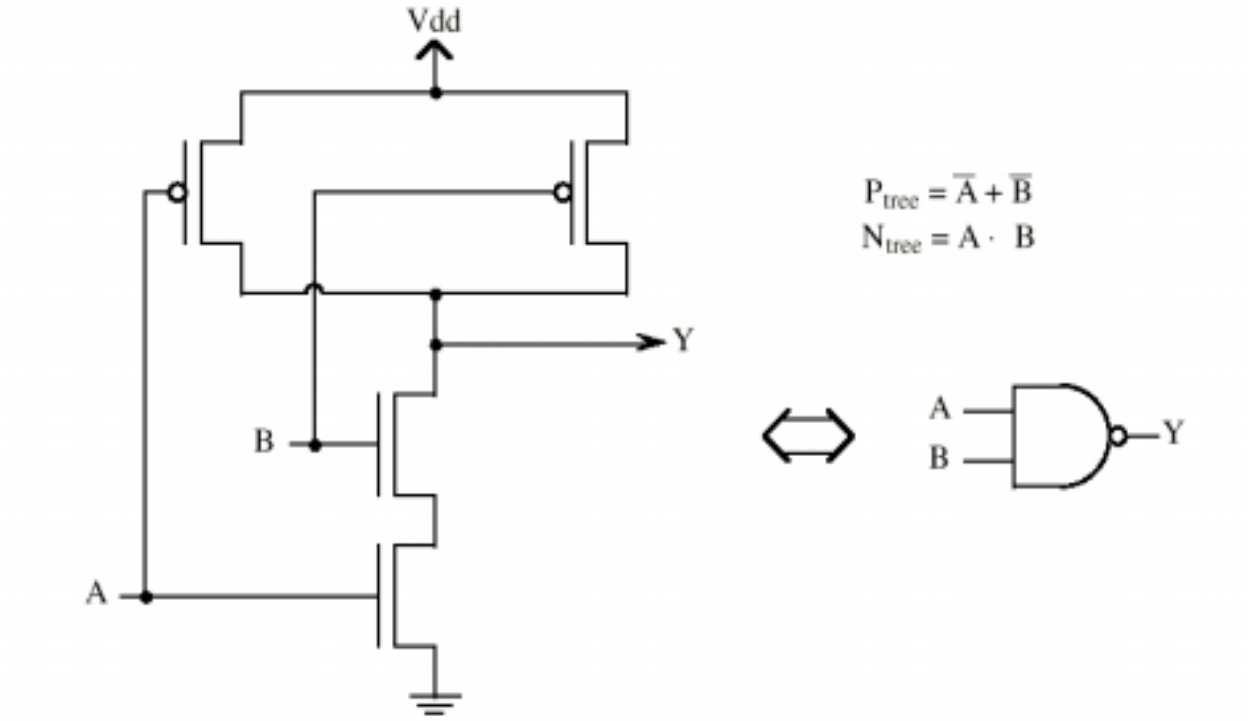
n-type transistor tree will provide "0" values of logic function

Truth Table (NAND):

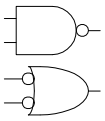
AB	
00	1
01	1
10	1
11	0

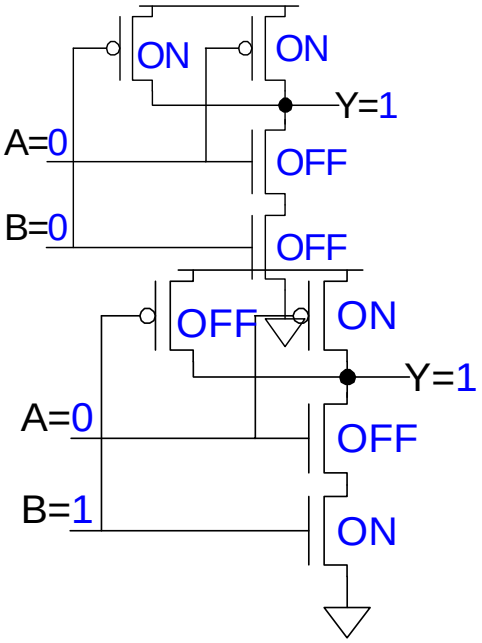
K-map (NAND):



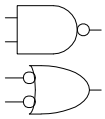


A	B	Y
0	0	1
0	1	
1	0	
1	1	

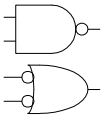


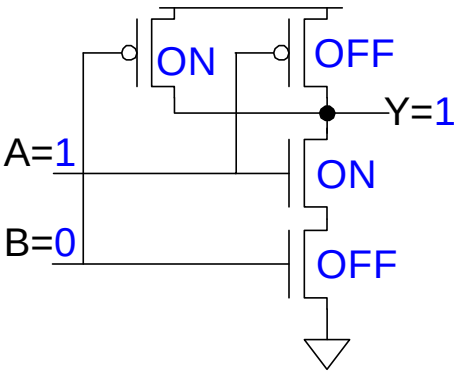


A	B	Y
0	0	1
0	1	1
1	0	
1	1	



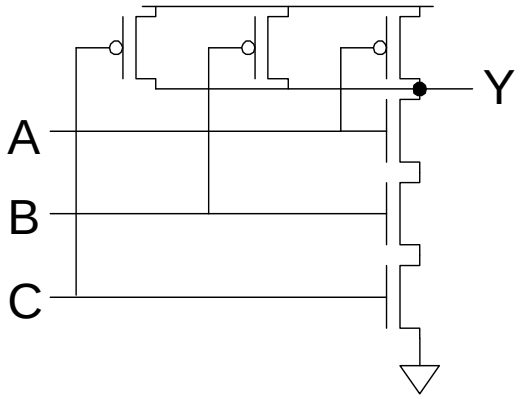
A	B	Y
0	0	1
0	1	1
1	0	1
1	1	0





NAND gate Design..

Y pulls low if ALL inputs are 1
Y pulls high if ANY input is 0



6.2 NOR gate Design..

NOR Gate Design

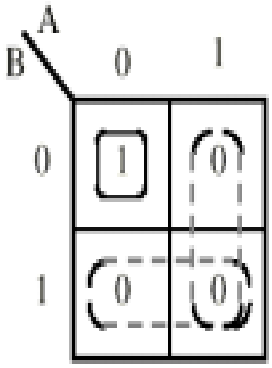
p-type transistor tree will provide "1" values of logic function

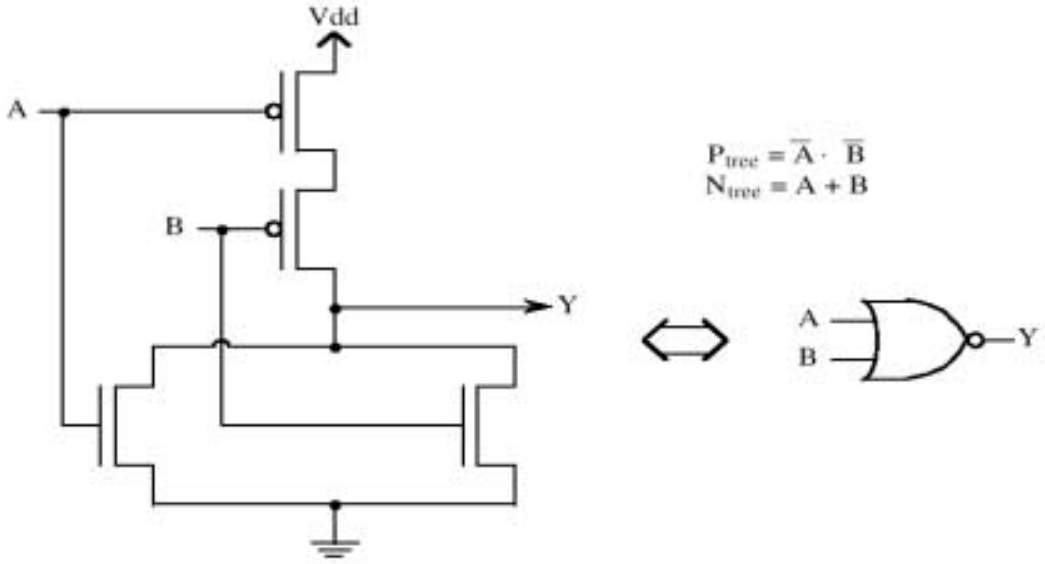
n-type transistor tree will provide "0" values of logic function

Truth Table:

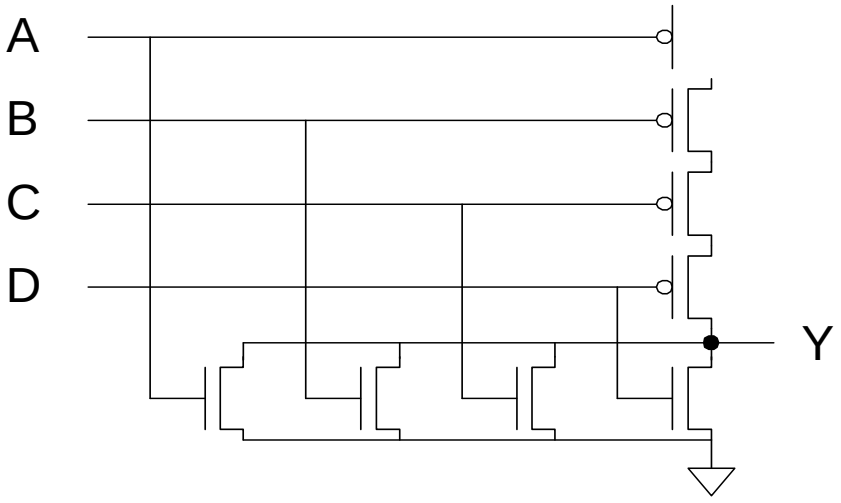
AB	
00	1
01	0
10	0
11	0

K-map:





4-input CMOS NOR gate



CMOS INVERTER

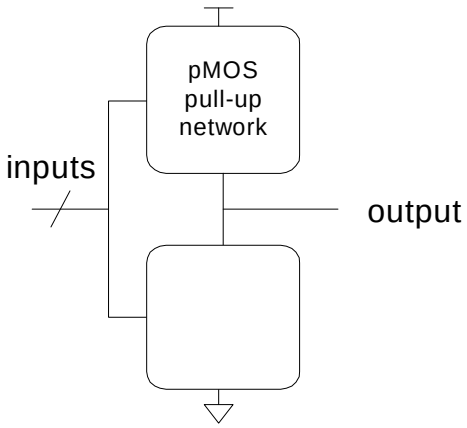
Note: Ideally there is no static power dissipation. When "I" is fully is *high* or fully *low*, no current path between Vdd and GND exists (the output is usually tied to the gate of another MOS transistor which has a very high input impedance).

Power is dissipated as "I" transistions from 0 →1 and 1→0 and a momentory current path exists between Vdd and GND. Power is also dissipated in the charging and discharging of gate capacitances.

6.3 CMOS Properties

Complementary CMOS logic gates
nMOS *pull-down network*
pMOS *pull-up netw* CMOS
Properties *ork*
a.k.a. static CMOS ,steady state
is reached to 0 or 1.(no dc path
from Vdd to gnd)

	Pull-up OFF	**Pull-up ON**
Pull-down OFF	**Z (float)**	**1**
Pull-down ON	**0**	**X (crowbar)**



nMOS
pull-down
network

- Complementary CMOS gates always produce 0 or 1
- Ex: NAND gate
- Series nMOS: Y=0 when both inputs are 1
- Thus Y=1 when either input is 0
- Requires parallel pMOS
- Rule of *Conduction Complements*

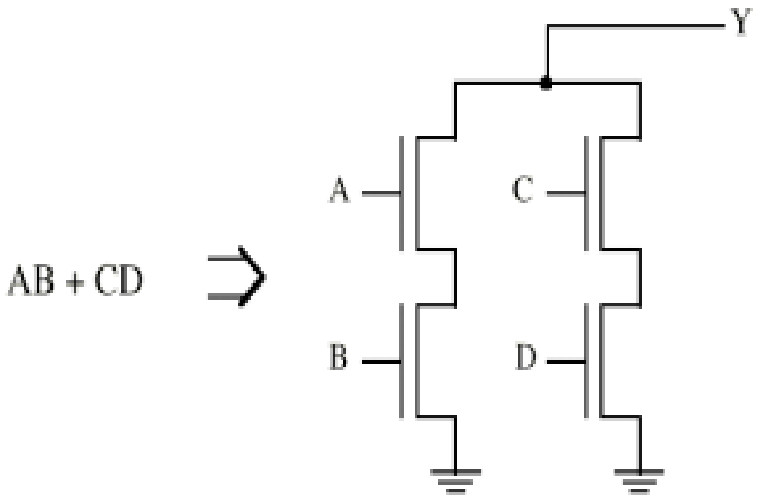
- Pull-up network is complement of pull-down
- Parallel -> series, series -> parallel
- Output signal strength is independent of input.-level restoring
- Restoring logic. Ouput signal strength is either V_{oh} (output high) or V_{ol}. (output low).
- Ratio less logic :output signal strength is independent of pMOS device size to nMOS size ratio.
- significant current only during the transition from one state to another and - hence power is conserved..
- Rise and fall transition times are of the same order,
- Very high levels of integration,
- High performance.

6.4 Complex gates..

$F = \overline{AB + CD} \quad \Rightarrow \quad N_{tree}$ will provide 0's, P_{tree} will provide 1's

0's of function F is $\overline{F}$, $\Rightarrow \overline{F} = \overline{\overline{AB + CD}} = AB + CD$

*n*MOS transistors need high true inputs, so it is desirable for all input variables to be high true, just as above.



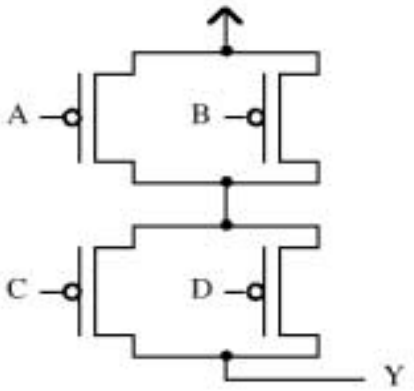
Likewise, a P_{tree} will provide 1's.

$F = \overline{AB + CD}$, need a form involving $\overline{A}$, $\overline{B}$, $\overline{C}$, $\overline{D}$

Apply DeMorgan's Theorem:

$F = \overline{AB} \cdot \overline{CD} = (\overline{A} + \overline{B}) \cdot (\overline{C} + \overline{D})$

Implementation ⇒

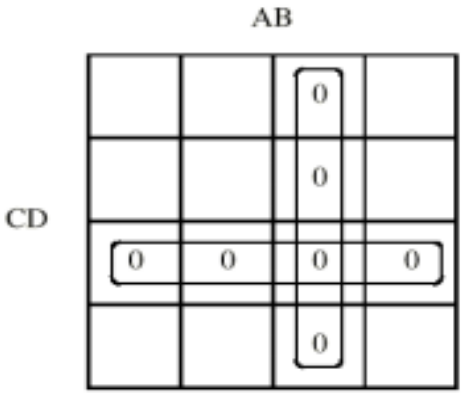


Can also use K-maps:

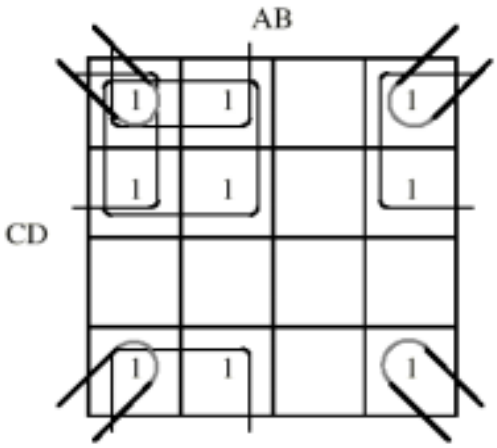
$F = \overline{AB + CD}$

CD \ AB				
	1	1	0	1
	1	1	0	1
	0	0	0	0
	1	1	0	1

For N_{tree}, minimize 0's; for P_{tree}, minimize 1's



$$N_{tree} = AB + CD$$

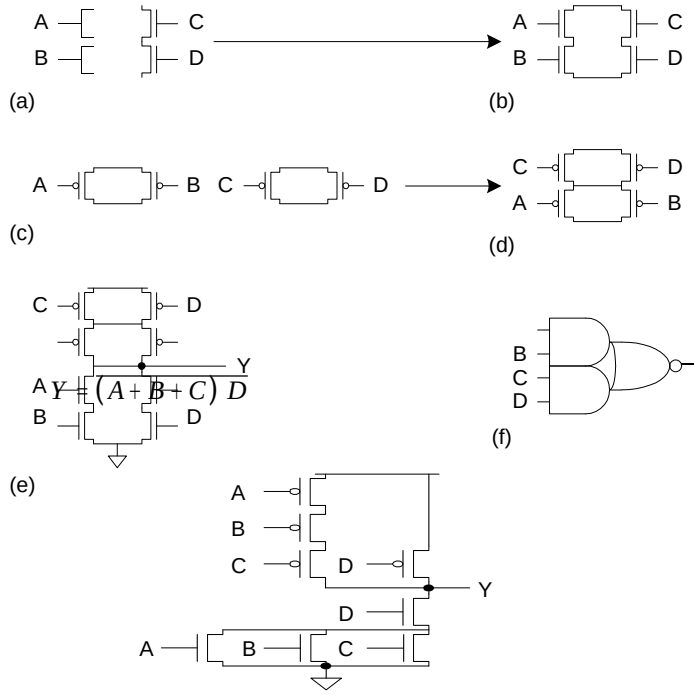


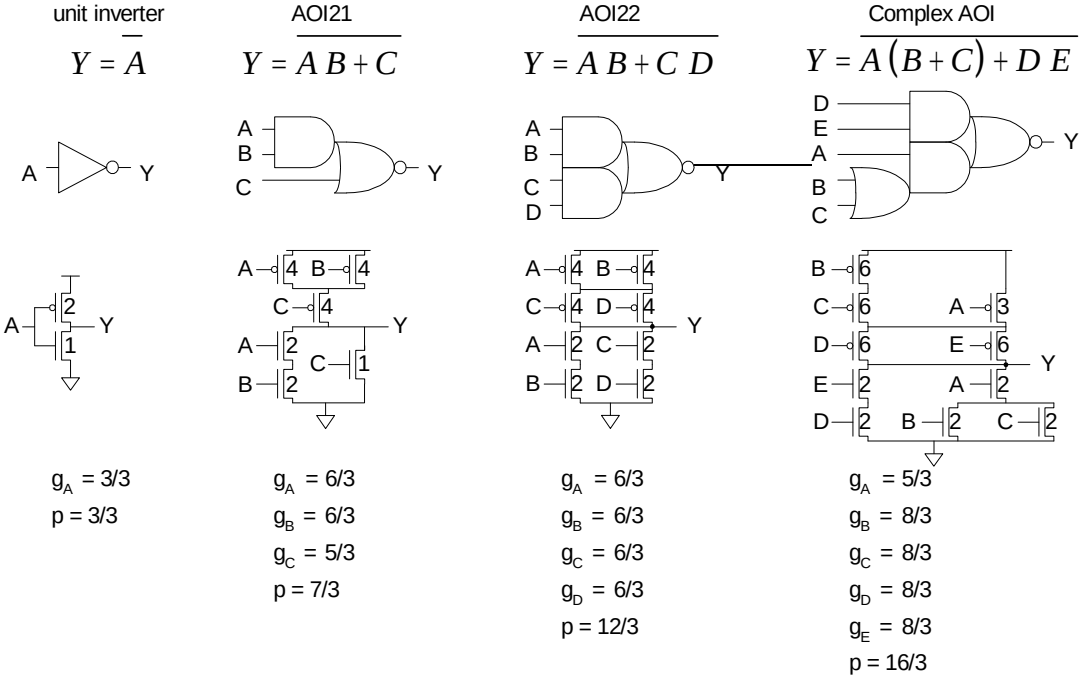
$$P_{tree} = \overline{A}\cdot\overline{C} + \overline{A}\cdot\overline{D} + \overline{B}\cdot\overline{C} + \overline{B}\cdot\overline{D}$$
$$= \overline{A}\,(\overline{C} + \overline{D}) + \overline{B}\,(\overline{C} + \overline{D})$$
$$= (\overline{A} + \overline{B}) \cdot (\overline{C} + \overline{D})$$

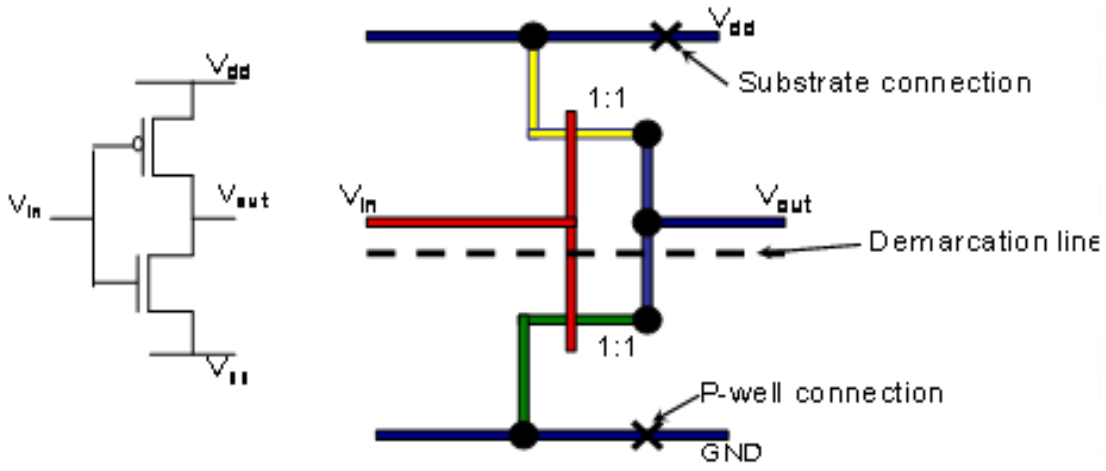
6.5 Complex gates AOI..

Compound gates can do any inverting function

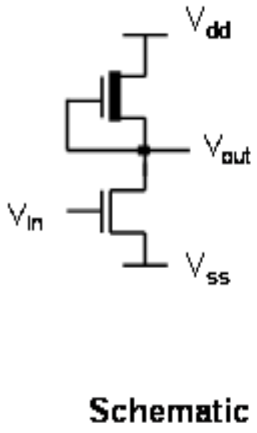
$Y = \overline{A\,B + C\,D}$ (AND-AND-OR-INVERT, AOI22)



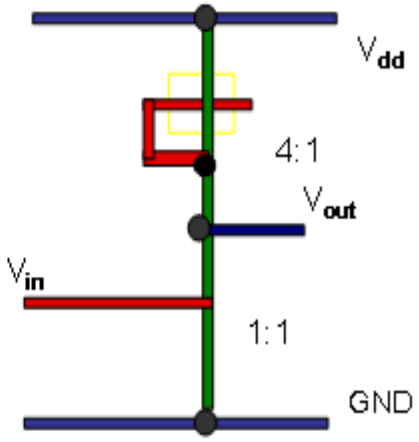




6.7 Restoring logic CMOS Variants: nMOS Inverter-stick diagram



Schematic



Stick diagram

- Basic inverter circuit: load replaced by depletion mode transistor
- With no current drawn from output, the current I_{ds} for both transistor must be same.
- For the depletion mode transistor, gate is connected to the source so it is always on and only the characteristic curve $V_{gs}=0$ is relevant.
- Depletion mode is called pull-up and the enhancement mode device pull-down.

- Obtain the transfer characteristics.
- As V_{in} exceeds the p.d. threshold voltage current begins to flow, V_{out} thus decreases and further increase will cause p.d transistor to come out of saturation and become resistive.
- p.u transistor is initially resistive as the p.d is turned on.
- Point at which $V_{out} = V_{in}$ is denoted as V_{inv}
- Can be shifted by variation of the ratio of pull-up to pull-down resistances $-Z_{p.u} / Z_{p.d}$
- Z- ratio of channel length to width for each transistor

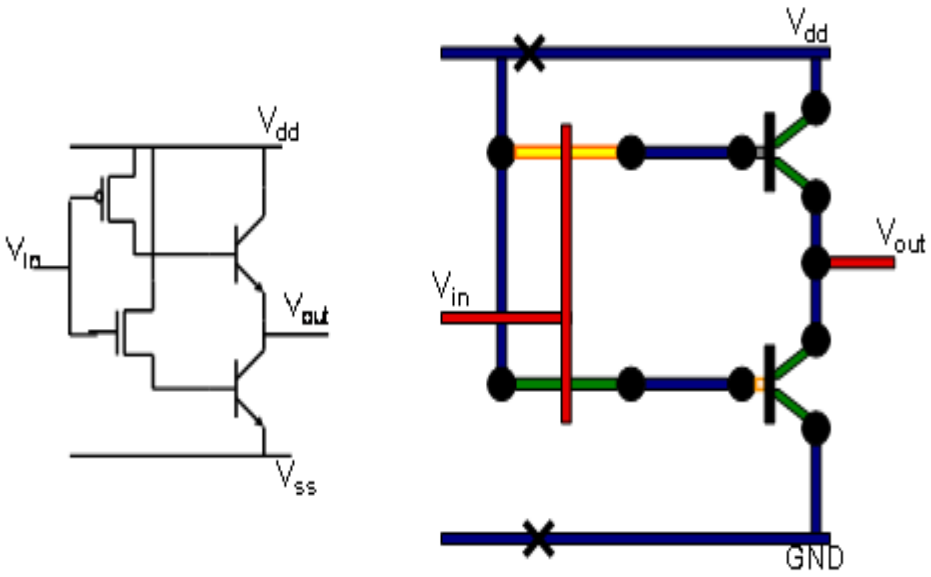
For 8:1 nMOS Inverter

$Z_{p.u.} = L_{p.u.} / W_{p.u} = 8$
$R_{p.u} = Z_{p.u.} * R_s = 80K$
similarly
$R_{p.d} = Z_{p.d} * R_s = 10K$
Power dissipation(on) $P_d = V^2 / R_{p.u} + R_{p.d} = 0.28mV$
Input capacitance = $1 C_g$

For 4:1 nMOS Inverter

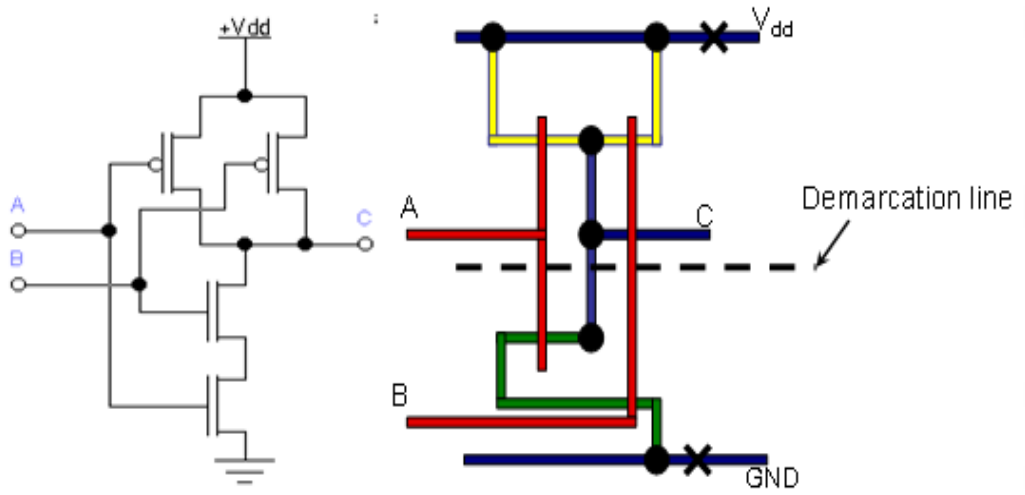
$Z_{p.u.} = L_{p.u.} / W_{p.u} = 4$
$R_{p.u} = Z_{p.u.} * R_s = 40K$
similarly
$R_{p.d} = Z_{p.d} * R_s = 5K$
Power dissipation(on) $P_d = V^2 / R_{p.u} + R_{p.d} = 0.56mV$
Input capacitance = $2C_g$

6.8Restoring logic CMOS Variants: BiCMOS Inverter-stick diagram

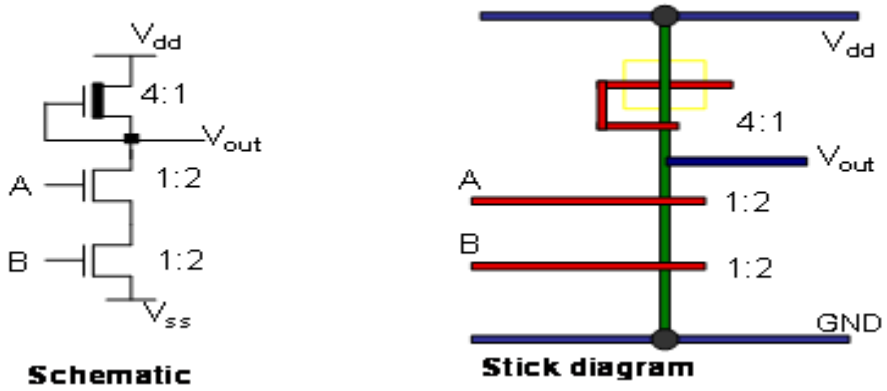


- A known deficiency of MOS technology is its limited load driving capabilities (due to limited current sourcing and sinking abilities of pMOS and nMOS transistors.)
- Output logic levels good-close to rail voltages
- High input impedance
- Low output impedance
- High drive capability but occupies a relatively small area.
- High noise margin
- Bipolar transistors have
    - higher gain
    - better noise characteristics
    - better high frequency characteristics
- BiCMOS gates can be an efficient way of speeding up VLSI circuits
- CMOS fabrication process can be extended for BiCMOS
- Example Applications
  CMOS- Logic
  BiCMOS- I/O and driver circuits
  ECL- critical high speed parts of the system

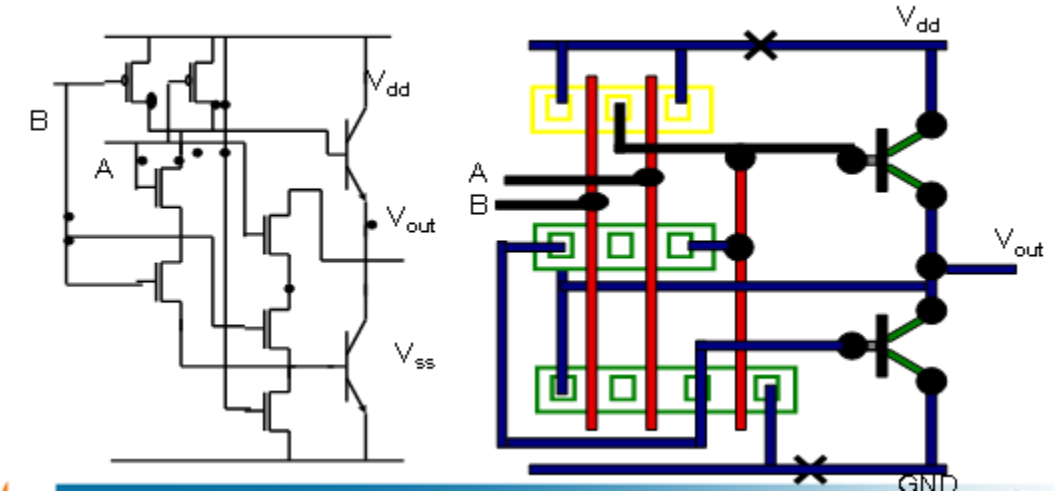
6.9 Circuit Families : Restoring logic CMOS NAND gate



6.10 Restoring logic CMOS Variants:_nMOS NAND gate



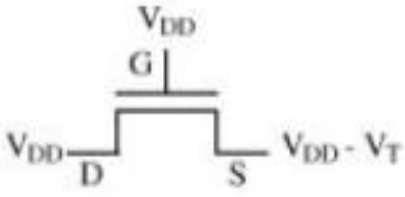
6.11 Restoring logic CMOS Variants: BiCMOS NAND gate



- For nMOS Nand-gate, the ratio between pull-up and sum of all pull-downs must be 4:1.
- nMOS Nand-gate area requirements are considerably greater than corresponding nMOS inverter
- nMOS Nand-gate delay is equal to number of input times inverter delay.
- Hence nMOS Nand-gates are used very rarely
- CMOS Nand-gate has no such restrictions
- BiCMOS gate is more complex and has larger fan-out.

7.Circuit Families :Switch logic: Pass Transistor

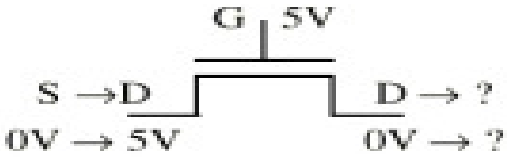
Why? *n*MOS switches cannot pass a logic "1" without a threshold voltage (V_T) drop.



where V_T = 0.7V to 1.0V (i.e., threshold voltage will vary)

output voltage = 4.3V to 4.0V, a *weak* "1"

The *n*MOS transistor will stop conducting if $V_{GS} < V_T$. Let V_T = 0.7V,



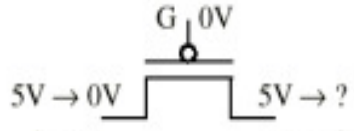
As source goes from 0V → 5V, V_{GS} goes from 5V → 0V.

When $V_S > 4.3V$, then $V_{GS} < V_T$, so switch stops conducting.

V_D left at 5V - V_T = 5V - 0.7V = 4.3V or Vdd - V_T.

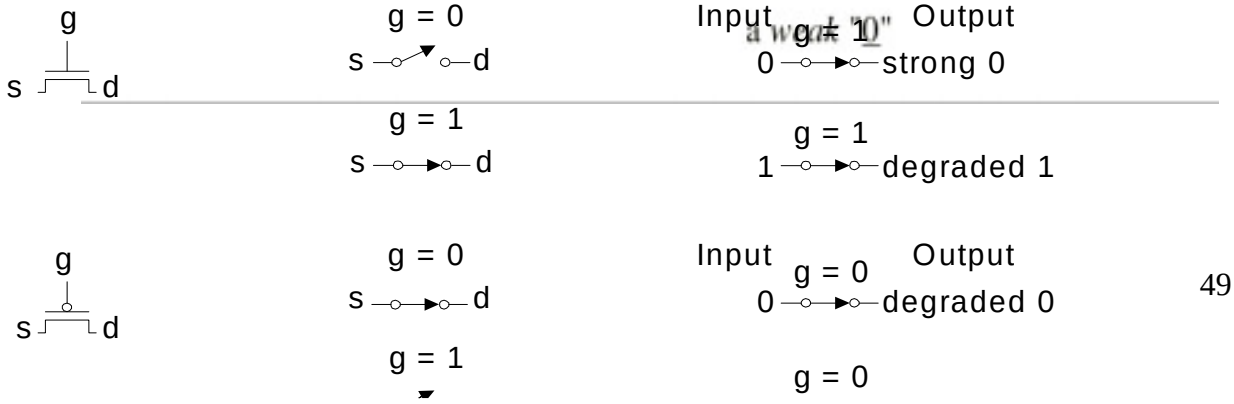
How will *p*MOS pass a "0"?

When $|V_{GS}| < |V_{Tp}|$, stop conducting



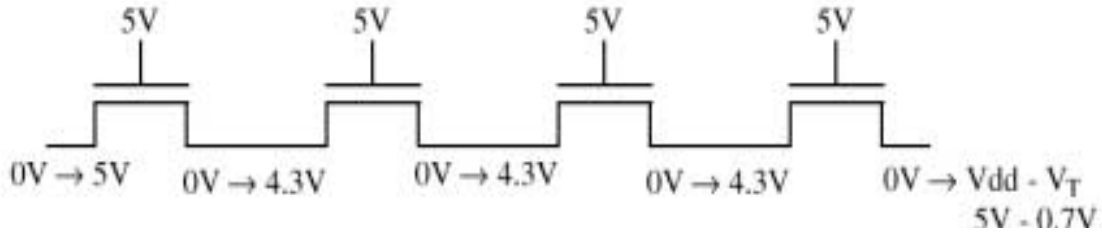
So when $|V_{GS}| < |-0.7V|$, V_D will go from 5V → **0.7V**,

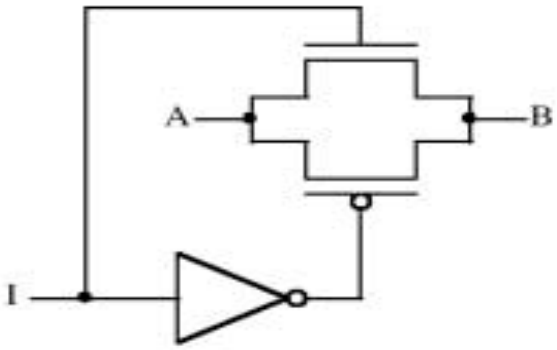
7.1 Switch logic: Pass Transistor



49

7.1 Switch logic: Pass Transistor-nMOS in series





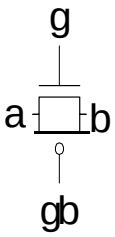
When I = 1,

B = strong 1, if A = 1;

B = strong 0, if A = 0

When I = 0, non-conducting

Pass transistors produce degraded outputs
Transmission gates pass both 0 and 1 well



Input Output

g = 0, gb = 1

a — b

g = 1, gb = 0

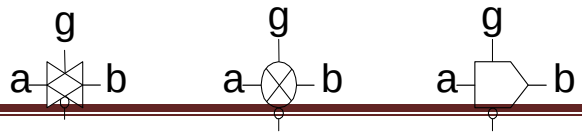
0 → strong 0

g = 1, gb = 0

a → b

g = 1, gb = 0

1 → strong 1

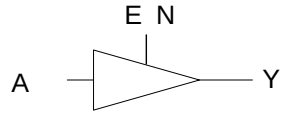


gb gb gb

EN	A	Y
0	0	
0	1	
1	0	
1	1	

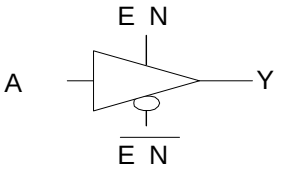
8 Structured Design-Tristate

- ***Tristate buffer*** **produces Z when not enabled**



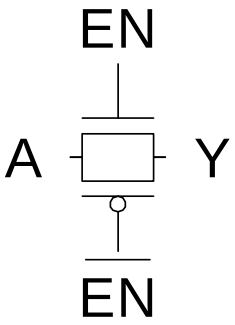
Tristate buffer **produces Z when not enabled**

EN	A	Y
0	0	Z
0	1	Z
1	0	0
1	1	1



8.1 Structured Design-Nonrestoring Tristate

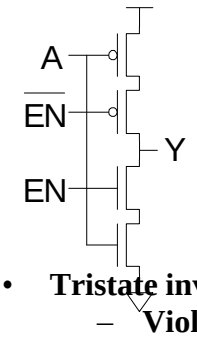
- Transmission gate acts as tristate buffer
  - two transistors
  - But *nonrestoring*
    - Noise on A is passed on to Y
  - + No V_t drop
  - - Requires inverted clock



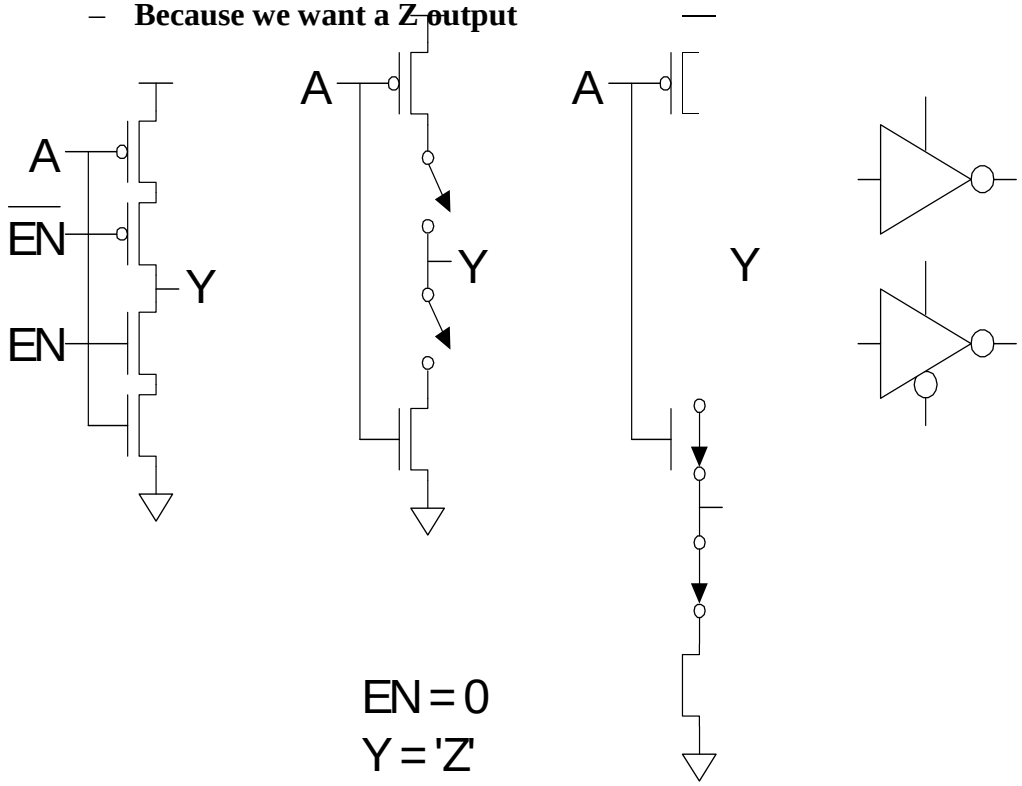
output
rule

8.3 Structured Design-Tristate Inverter

- Tristate inverter produces restored
  - Violates conduction complement
  - Because we want a Z output



- **Tristate inverter produces restored output**
  - **Violates conduction complement rule**
  - **Because we want a Z output**



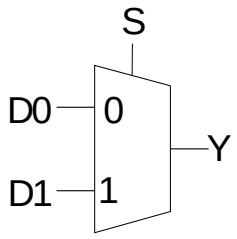
8.4 Structured Design-Multiplexers

- **2:1 *multiplexer* chooses between two inputs**

S	D1	D0	Y
0	X	0	0
0	X	1	1
1	0	X	0
1	1	X	1

S D1 D0 Y

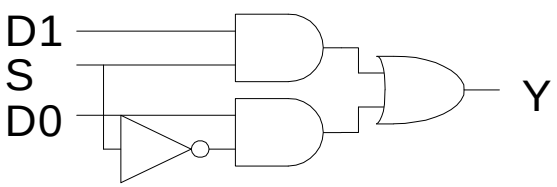
0	X	0	
0	X	1	
1	0	X	
1	1	X	

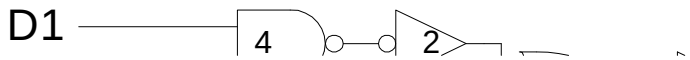


8. 5 Structured Design-Mux Design.. Gate-Level

$$Y = SD_1 + \overline{S}D_0 \text{ (too many transistors)}$$

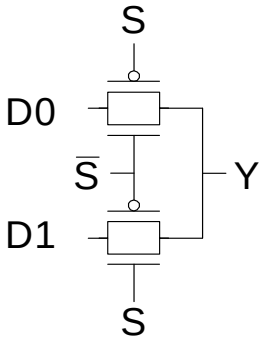
- **How many transistors are needed?**
- **How many transistors are needed? 20**




D1
4
2

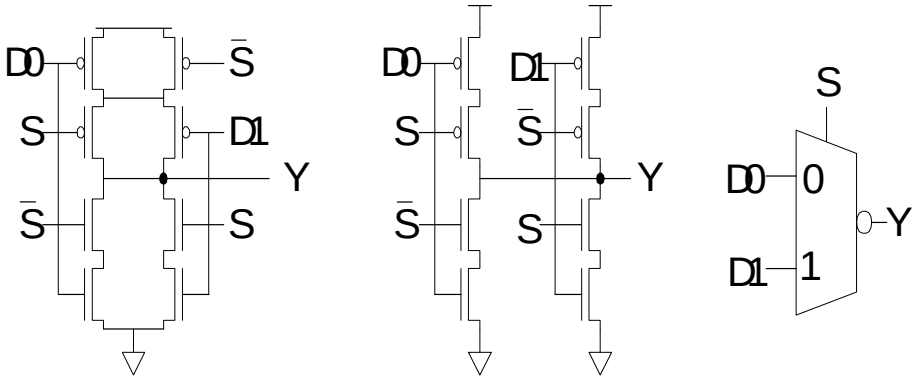
8.6 Structured Design-Mux Design-Transmission Gate

- Nonrestoring mux uses two transmission gates
  - Only 4 transistors



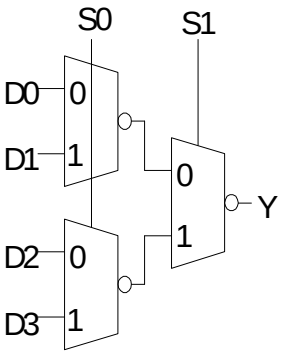
Inverting Mux

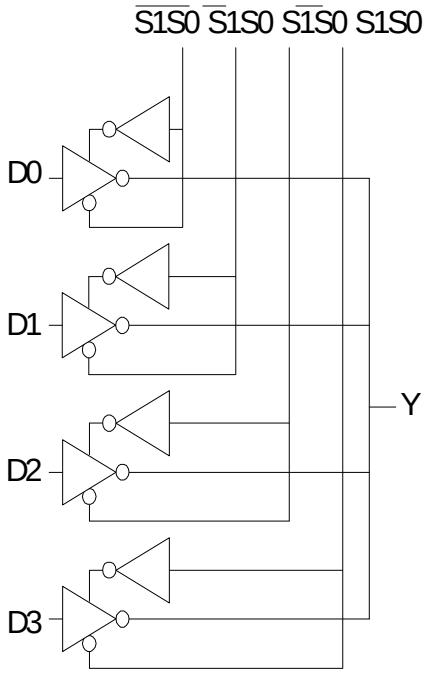
- Inverting multiplexer
  - Use compound AOI22
  - Or pair of tristate inverters
- **Noninverting multiplexer adds an inverter**



8.7 Design-4:1 Multiplexer

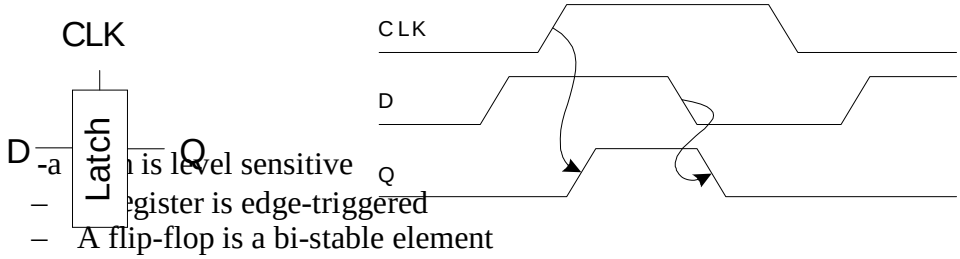
- 4:1 mux chooses one of 4 inputs using two selects
  Two levels of 2:1 muxes
  Or four tristates





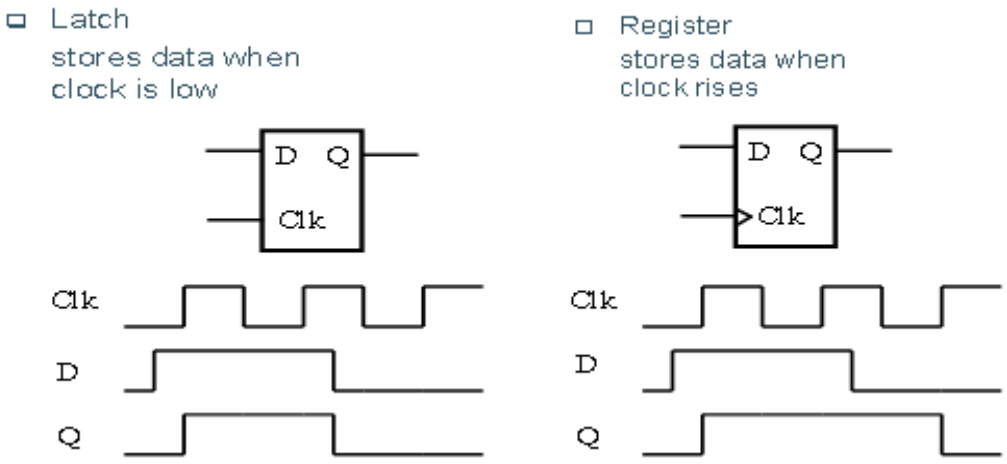
9 Structured Design-D Latch

- When CLK = 1, latch is *transparent*
  - D flows through to Q like a buffer
- When CLK = 0, the latch is *opaque*
  - Q holds its old value independent of D
- a.k.a. *transparent latch* or *level-sensitive latch*



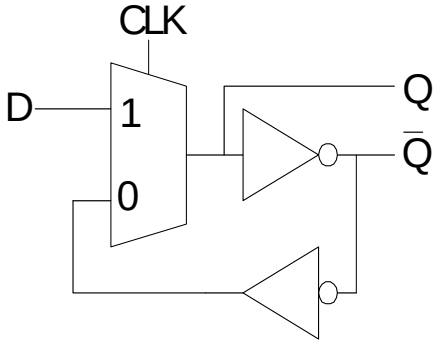
- -a latch is level sensitive
  - A register is edge-triggered
  - A flip-flop is a bi-stable element

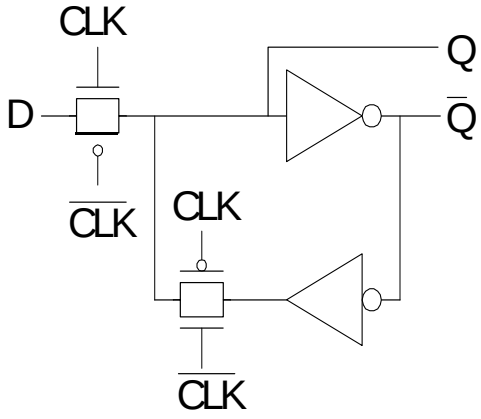
–



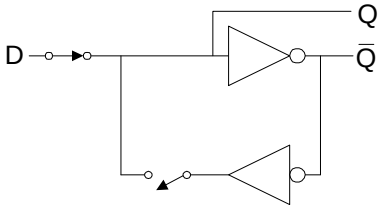
9.1 D Latch Design

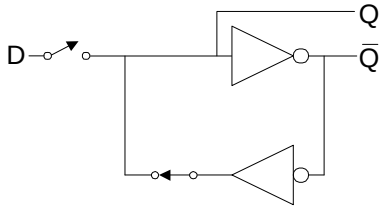
- Multiplexer chooses D or old Q





9.2 D Latch Operation





CLK = 1

CLK = 0

UNIT - 7

SUBSYSTEM DESIGN PROCESSES: Some general considerations, an Illustration of design process, Observations **4 Hours**

UNIT - 8

ILLUSTRATION OF THE DESIGN PROCESS: Observation on the design process, Regularity Design of an ALU subsystem. Design of 4-bit adder, implementing ALU functions. **4 Hours**

Objectives: At the end of this unit we will be able to understand

- Design consideration, problem and solution
- Design processes
- Basic digital processor structure
- Datapath
- Bus Architecture
- Design 4 – bit shifter
- Design of ALU subsystem
- 4 – bit Adder

General Considerations

Lower unit cost
Higher reliability
Lower power dissipation, lower weight and lower volume
Better performance
Enhanced repeatability
Possibility of reduced design/development periods

Some Problems

1. How to design complex systems in a reasonable time & with reasonable effort.
2. The nature of architectures best suited to take full advantage of VLSI and the technology
3. The testability of large/complex systems once implemented on silicon

Some Solution

Problem 1 & 3 are greatly reduced if two aspects of standard practices are accepted.

1. a) Top-down design approach with adequate CAD tools to do the job

   b) Partitioning the system sensibly

   c) Aiming for simple interconnections

   d) High regularity within subsystem

   e) Generate and then verify each section of the design
2. Devote significant portion of total chip area to test and diagnostic facility
3. Select architectures that allow design objectives and high regularity in realization

Illustration of design processes

1. Structured design begins with the concept of hierarchy
2. It is possible to divide any complex function into less complex subfunctions that is up to leaf cells
3. Process is known as top-down design
4. As a systems complexity increases, its organization changes as different factors become relevant to its creation
5. Coupling can be used as a measure of how much submodels interact
6. It is crucial that components interacting with high frequency be physically proximate, since one may pay severe penalties for long, h igh-bandwidth interconnects
7. Concurrency should be exploited – it is desirable that all gates on the chip do useful work most of the time
8. Because technology changes so fast, the adaptation to a new process must occur in a short time.

Hence representing a design several approaches are possible. They are:

- Conventional circuit symbols
- Logic symbols
- Stick diagram
- Any mixture of logic symbols and stick diagram that is convenient at a stage
- Mask layouts
- Architectural block diagrams and floor plans

General arrangements of a 4 – bit arithmetic processor

The basic architecture of digital processor structure is as shown below in figure 6.1. Here the design of datapath is only considered.

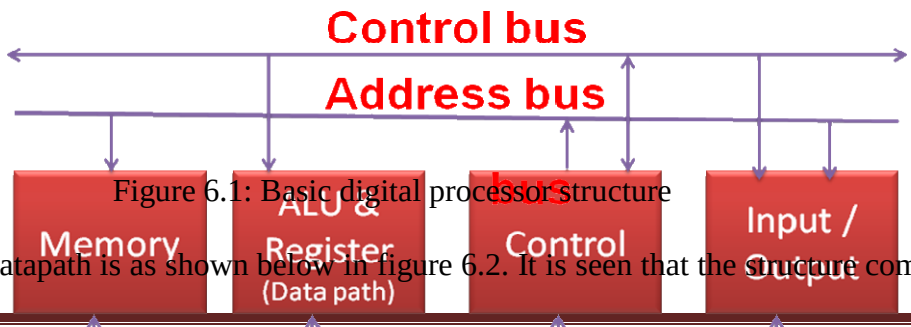


Figure 6.1: Basic digital processor structure

Datapath is as shown below in figure 6.2. It is seen that the structure comprises of

a unit which processes data applied at one port and presents its output at a second port.

Alternatively, the two data ports may be combined as a single bidirectional port if storage facilities exist in the datapath. Control over the functions to be performed is effected by control signals as shown.

Figure 6.2: Communication strategy for the datapath

Datapath can be decomposed into blocks showing the main subunits as in figure 3. In doing so it is useful to anticipate a possible floor plan to show the planned relative decomposition of the subunits on the chip and hence on the mask layouts.

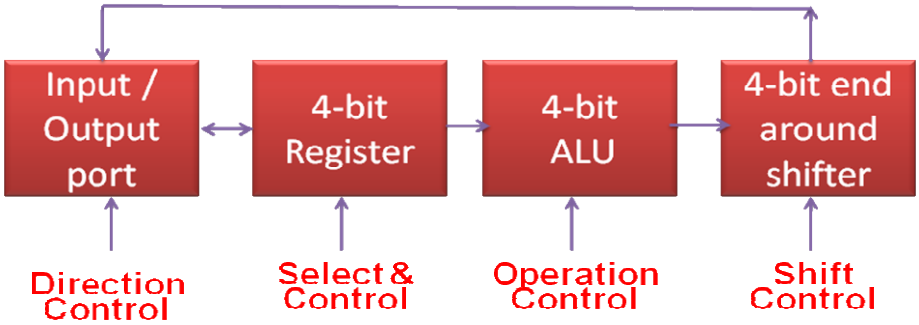


Figure 6.3: Subunits and basic interconnection for datapath

Nature of the bus architecture linking the subunits is discussed below. Some of the possibilities are:

One bus architecture:

Figure 6.4: One bus architecture

Sequence:

1. 1st operand from registers to ALU. Operand is stored there.
2. 2nd operand from register to ALU and added.
3. Result is passed through shifter and stored in the register

Two bus architecture:

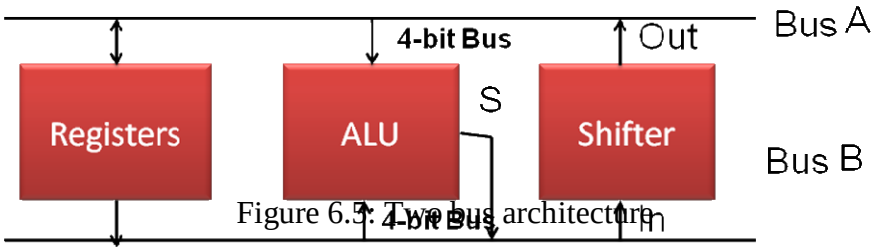


Figure 6.5: Two bus architecture

Sequence:

1. Two operands (A & B) are sent from register(s) to ALU & are operated upon, result S in ALU.
2. Result is passed through the shifter & stored in registers.

Three bus architecture:

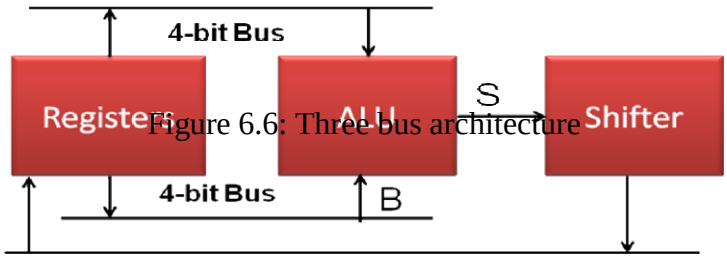


Figure 6.6: Three bus architecture

Sequence:

Two operands (A & B) are sent from registers, operated upon, and shifted result (S) returned to another register, all in same clock period.

In pursuing this design exercise, it was decided to implement the structure with a 2 – bus architecture. A tentative floor plan of the proposed design which includes some form of interface to the parent system data bus is shown in figure 6.7.

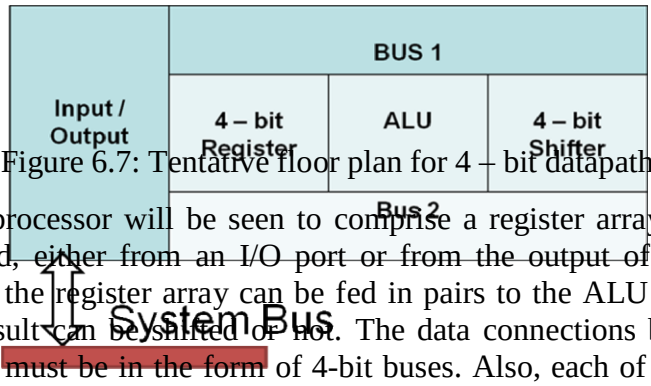


Figure 6.7: Tentative floor plan for 4 – bit datapath

The proposed processor will be seen to comprise a register array in which 4-bit numbers can be stored, either from an I/O port or from the output of the ALU via a shifter. Numbers from the register array can be fed in pairs to the ALU to be added (or subtracted) and the result can be shifted or not. The data connections between the I/O port, ALU, and shifter must be in the form of 4-bit buses. Also, each of the blocks must be suitably connected to control lines so that its function may be defined for any of a range of possible operations.

During the design process, and in particular when defining the interconnection strategy and designing the stick diagrams, care must be taken in allocating the layers to the various data or control paths. Points to be noted:

Metal can cross poly or diffusion
Poly crossing diffusion form a transistor
Whenever lines touch on the same level an interconnection is formed
Simple contacts can be used to join diffusion or poly to metal.
Buried contacts or a butting contacts can be used to join diffusion and poly

Some processes use 2nd metal

1st and 2nd metal layers may be joined using a via
Each layer has particular electrical properties which must be taken into account
For CMOS layouts, p-and n-diffusion wires must not directly join each other
Nor may they cross either a p-well or an n-well boundary

Design of a 4-bit shifter

Any general purpose n-bit shifter should be able to shift incoming data by up to n – 1 place in a right-shift or left-shift direction. Further specifying that all shifts should be on an end-around basis, so that any bit shifted out at one end of a data word will be shifted in at the other end of the word, then the problem of right shift or left shift is greatly eased. It can be analyzed that for a 4-bit word, that a 1-bit shift right is equivalent to a 3-bit shift left and a 2-bit shift right is equivalent to a 2-bit left etc. Hence, the design of either shift right or left can be done. Here the design is of shift right by 0, 1, 2, or 3 places. The shifter must have:

- input from a four line parallel data bus
- four output lines for the shifted data
- means of transferring input data to output lines with any shift from 0 to 3 bits

Consider a direct MOS switch implementation of a 4 X 4 crossbar switches shown in

figure 6.8. The arrangement is general and may be expanded to accommodate n-bit inputs/outputs. In this arrangement any input can be connected to any or all the outputs. Furthermore, 16 control signals ($sw_{00} - sw_{15}$), one for each transistor switch, must be provided to drive the crossbar switch, and such complexity is highly undesirable.

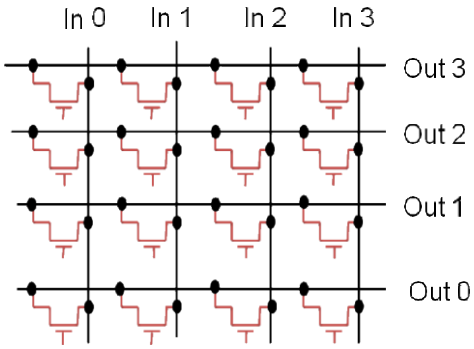


Figure 6.8: 4 X 4 crossbar switch

An adaptation of this arrangement recognizes the fact that we couple the switch gates together in groups of four and also form four separate groups corresponding to shifts of zero, one, two and three bits. The resulting arrangement is known as a barrel shifter and a 4 X 4 barrel shifter circuit diagram is as shown in the figure 6.9.

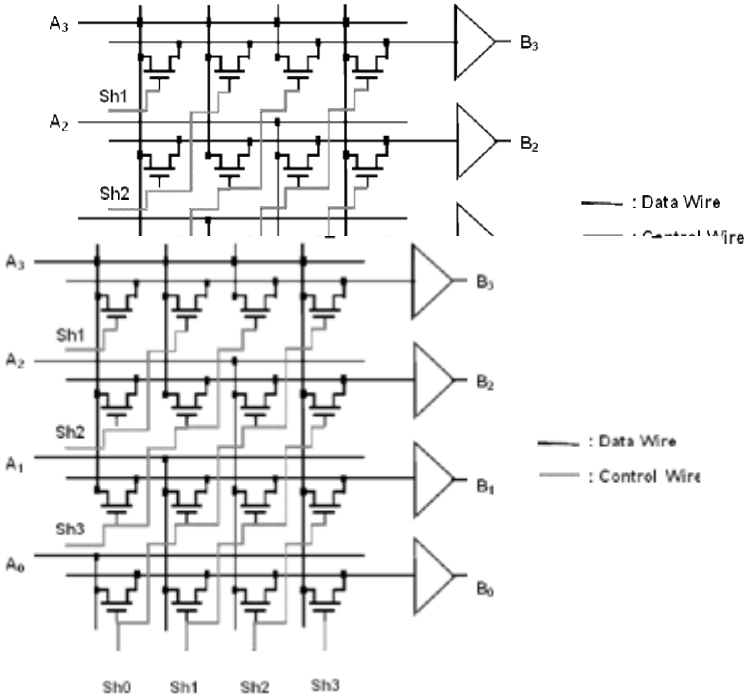


figure 6.9: 4 X 4 barrel shifter

The interbus switches have their gate inputs connected in a staircase fashion in groups of four and there are now four shift control inputs which must be mutually exclusive in the active state. CMOS transmission gates may be used in place of the simple pass transistor switches if appropriate. Barrel shifter connects the input lines representing a word to a group of output lines with the required shift determined by its control inputs (sh0, sh1, sh2, sh3). Control inputs also determine the direction of the shift. If input word has n – bits and shifts from 0 to n-1 bit positions are to be implemented.

To summaries the design steps

- Set out the specifications

- Partition the architecture into subsystems
- Set a tentative floor plan
- Determine the interconnects
- Choose layers for the bus & control lines
- Conceive a regular architecture
- Develop stick diagram
- Produce mask layouts for standard cell
- Cascade & replicate standard cells as required to complete the design

Design of an ALU subsystem

Having designed the shifter, we shall design another subsystem of the 4-bit data path. An appropriate choice is ALU as shown in the figure 6.10 below.

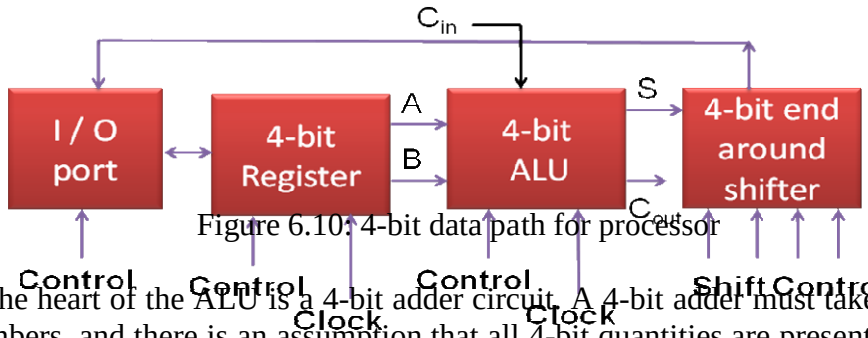


Figure 6.10: 4-bit data path for processor

The heart of the ALU is a 4-bit adder circuit. A 4-bit adder must take sum of two 4-bit numbers, and there is an assumption that all 4-bit quantities are presented in parallel form and that the shifter circuit is designed to accept and shift a 4-bit parallel sum from the ALU. The sum is to be stored in parallel at the output of the adder from where it is fed through the shifter and back to the register array. Therefore, a single 4-bit data bus is needed from the adder to the shifter and another 4-bit bus is required from the shifted output back to the register array. Hence, for an adder two 4-bit parallel numbers are fed on two 4-bit buses. The clock signal is also required to the adder, during which the inputs are given and sum is generated. The shifter is unclocked but must be connected to four shift control lines.

Design of a 4-bit adder:

The truth table of binary adder is as shown in table 6.1

As seen from the table any column k there will be three inputs namely A_k , B_k as present input number and C_{k-1} as the previous carry. It can also be seen that there are two outputs sum S_k and carry C_k.

From the table one form of the equation is:

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

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

Where

Half sum $H_k = A_k' B_k + A_k B_k'$

Adder element requirements

Table 6.1 reveals that the adder requirement may be stated as:

If $A_k = B_k$ then $S_k = C_{k-1}$

Else $S_k = C_{k-1}'$

And for the carry C_k

If $A_k = B_k$ then $C_k = A_k = B_k$

Else $C_k = C_{k-1}$

Thus the standard adder element for 1-bit is as shown in the figure 6.11.

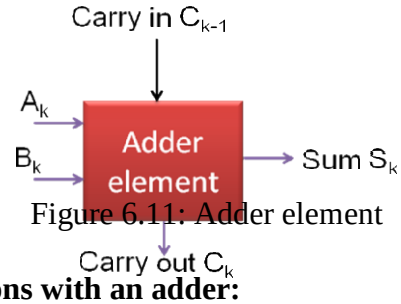


Figure 6.11: Adder element

Implementing ALU functions with an adder:

An ALU must be able to add and subtract two binary numbers, perform logical operations such as And, Or and Equality (Ex-or) functions. Subtraction can be performed by taking 2's complement of the negative number and perform the further addition. It is desirable to keep the architecture as simple as possible, and also see that the adder performs the logical operations also. Hence let us examine the possibility.

The adder equations are:

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

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

Where

Half sum $H_k = A_k' B_k + A_k B_k'$

Let us consider the sum output, if the previous carry is at logical 0, then

$S_k = H_k . 1 + H_k' . 0$

$S_k = H_k = A_k'B_k + A_k B_k'$ – An Ex-or operation

Now, if C_{k-1} is logically 1, then

$S_k = H_k . 0 + H_k' . 1$

$S_k = H_k'$ – An Ex-Nor operation

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

$C_k = A_kB_k + H_k . 0$

$C_k = A_kB_k$ - An And operation

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

$C_k = A_kB_k + H_k . 1$

On solving $C_k = A_k + B_k$ - An Or operation

The adder element implementing both the arithmetic and logical functions can be implemented as shown in the figure 6.12.

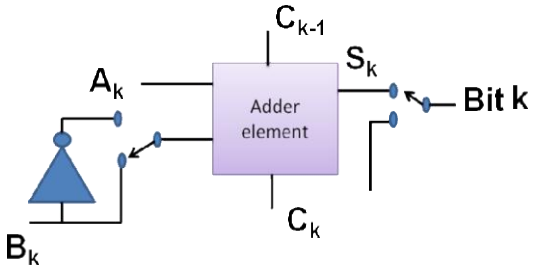


Figure 6.12: 1-bit adder element

The above can be cascaded to form 4-bit ALU.

A further consideration of adders

Generation:

This principle of generation allows the system to take advantage of the occurrences "$a_k=b_k$". In both cases ($a_k=1$ or $a_k=0$) the carry bit will be known.

Propagation:

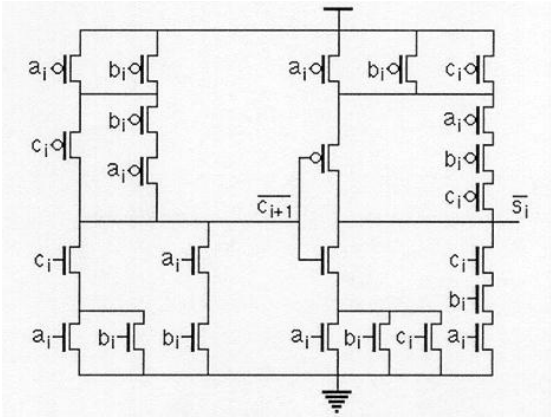
If we are able to localize a chain of bits $a_k\ a_{k+1}...a_{k+p}$ and $b_k\ b_{k+1}...b_{k+p}$ for which a_k not equal to b_k for k in [k,k+p], then the output carry bit of this chain will be equal to the input carry bit of the chain.

These remarks constitute the principle of generation and propagation used to speed the addition of two numbers.

All adders which use this principle calculate in a first stage.

$p_k = a_k$ XOR b_k

$g_k = a_k\ b_k$



Manchester carry – chain

This implementation can be very performant (20 transistors) depending on the way the XOR function is built. The carry propagation of the carry is controlled by the output of the XOR gate. The generation of the carry is directly made by the function at the bottom. When both input signals are 1, then the inverse output carry is 0.

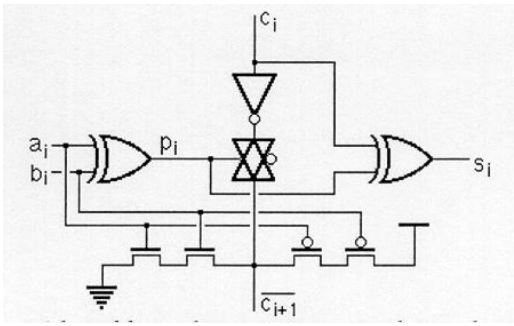


Figure-6.12: An adder with propagation signal controlling the pass-gate

In the schematic of Figure 6.12, the carry passes through a complete transmission gate. If the carry path is precharged to VDD, the transmission gate is then reduced to a simple NMOS transistor. In the same way the PMOS transistors of the carry generation is removed. One gets a Manchester cell.

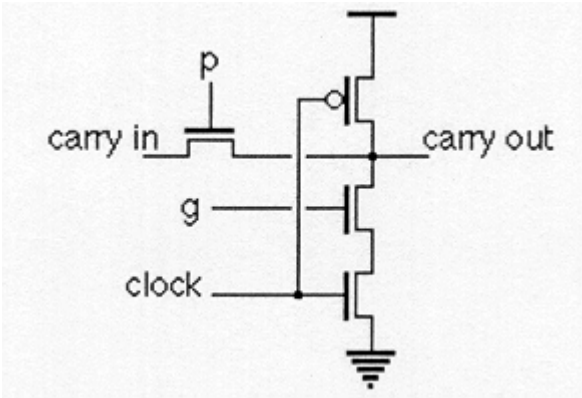


Figure-6.13: The Manchester cell

The Manchester cell is very fast, but a large set of such cascaded cells would be slow. This is due to the distributed RC effect and the body effect making the propagation time grow with the square of the number of cells. Practically, an inverter is added every four cells, like in Figure 6.14.

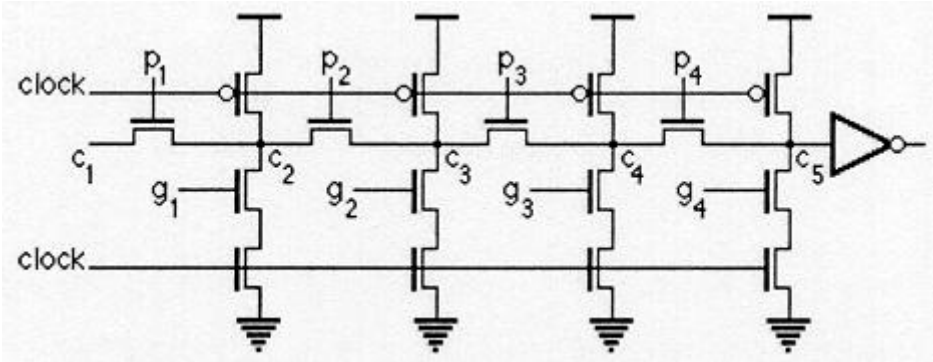


Figure-6.14**:** The Manchester carry cell

Adder Enhancement techniques

The operands of addition are the addend and the augend. The addend is added to the augend to form the sum. In most computers, the augmented operand (the augend) is replaced by the sum, whereas the addend is unchanged. High speed adders are not only for addition but also for subtraction, multiplication and division. The speed of a digital processor depends heavily on the speed of adders. The adders add vectors of bits and the principal problem is to speed- up the carry signal. A traditional and non optimized four bit adder can be made by the use of the generic one-bit adder cell connected one to the other. It is the ripple carry adder. In this case, the sum resulting at each stage need to wait for the incoming carry signal to perform the sum operation. The carry propagation can be speed-up in two ways. The first –and most obvious– way is to use a faster logic circuit technology. The second way is to generate carries by means of forecasting logic that does not rely on the carry signal being rippled from stage to stage of the adder.

The Carry-Skip Adder

Depending on the position at which a carry signal has been generated, the propagation time can be variable. In the best case, when there is no carry generation, the addition time will only take into account the time to propagate the carry signal. Figure 6.15 is an example illustrating a carry signal generated twice, with the input carry being equal to 0. In this case three simultaneous carry propagations occur. The longest is the second, which takes 7 cell delays (it starts at the 4th position and ends at the 11th position). So the addition time of these two numbers with this 16-bits Ripple Carry Adder is 7.k + k', where k is the delay cell and k' is the time needed to compute the 11th sum bit using the 11th carry-in.

With a Ripple Carry Adder, if the input bits Ai and Bi are different for all position i, then the carry signal is propagated at all positions (thus never generated), and the addition is completed when the carry signal has propagated through the whole adder. In this case, the Ripple Carry Adder is as slow as it is large. Actually, Ripple Carry Adders are fast only for some configurations of the input words, where carry signals are generated at some positions.

Carry Skip Adders take advantage both of the generation or the propagation of the carry signal. They are divided into blocks, where a special circuit detects quickly if all the bits to be added are different (Pi = 1 in all the block). The signal produced by this circuit will be called block propagation signal. If the carry is propagated at all positions in the block, then the carry signal entering into the block can directly bypass it and so be transmitted through a multiplexer to the next block. As soon as the carry signal is transmitted to a block, it starts to propagate through the block, as if it had been generated at the beginning of the block. Figure 6.16 shows the structure of a 24-bits Carry Skip Adder, divided into 4 blocks.

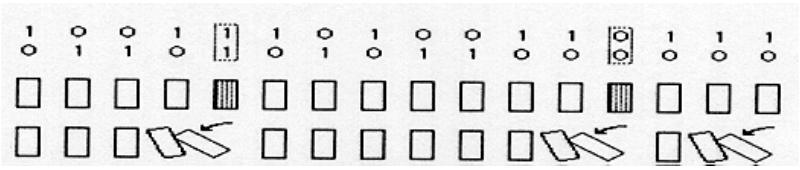


Figure 6.15: Example of Carry skip adder

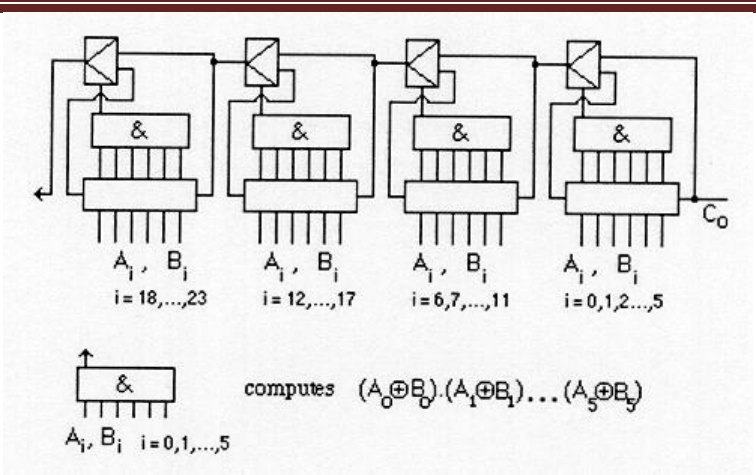


Figure-6.16: Block diagram of a carry skip adder

Baugh-Wooley Multiplier

This technique has been developed in order to design regular multipliers, suited for 2's-complement numbers.

Let us consider 2 numbers A and B:

$$\mathbf{A} = (\mathbf{a_{n-1}} \dots \mathbf{a_0}) = -\mathbf{a_{n-1}} \cdot 2^{\mathbf{n-1}} + \sum_{0}^{\mathbf{n-2}} \mathbf{a_i} \cdot 2^{\mathbf{i}}$$

$$\mathbf{B} = (\mathbf{b_{n-1}} \dots \mathbf{b_0}) = -\mathbf{b_{n-1}} \cdot 2^{\mathbf{n-1}} + \sum_{0}^{\mathbf{n-2}} \mathbf{b_i} \cdot 2^{\mathbf{i}}$$

The product A.B is given by the following equation:

$$\mathbf{A} \cdot \mathbf{B} = \mathbf{a_{n-1}} \cdot \mathbf{b_{n-1}} \cdot 2^{2\mathbf{n-2}} + \sum_{0}^{\mathbf{n-2}} \sum_{0}^{\mathbf{n-2}} \mathbf{a_i} \cdot \mathbf{b_i} \cdot 2^{\mathbf{i+j}} - \mathbf{a_{n-1}} \sum_{0}^{\mathbf{n-2}} \mathbf{b_i} \cdot 2^{\mathbf{n+i-1}} - \mathbf{b_{n-1}} \sum_{0}^{\mathbf{n-2}} \mathbf{a_i} \cdot 2^{\mathbf{n+i-1}}$$

We see that subtraction cells must be used. In order to use only adder cells, the negative terms may be rewritten as:

$$-a_{n-1}\sum_{0}^{n-2} b_i . 2^{i+n-1} = a_{n-1}.\left(-2^{2n-2} + 2^{n-1} + \sum_{0}^{n-2} \overline{b_i}. 2^{i+n-1}\right)$$

By this way, A.B becomes:

$$A.B = a_{n-1}.b_{n-1}.2^{2n-2} + \sum_{0}^{n-2}\sum_{0}^{n-2} a_i.b_j.2^{i+j}$$

$$+ b_{n-1}\left[-2^{2n-2} + 2^{n-1} + \sum_{0}^{n-2} \overline{a_i}.2^{i+n-1}\right]$$

$$+ a_{n-1}\left[-2^{2n-2} + 2^{n-1} + \sum_{0}^{n-2} \overline{b_i}.2^{i+n-1}\right]$$

The final equation is:

$$A.B = -2^{2n-1} + \left(\overline{a_{n-1}} + \overline{b_{n-1}} + a_{n-1}.b_{n-1}\right).2^{2n-2}$$

$$+\sum_{0}^{n-2}\sum_{0}^{n-2} a_i.b_j.2^{i+j} + \left(a_{n-1} + b_{n-1}\right).2^{n-1}$$

$$+\sum_{0}^{n-2} b_{n-1}.\overline{a_i}.2^{i+n-1} + \sum_{0}^{n-2} a_{n-1}.\overline{b_i}.2^{i+n-1}$$

because:

$$-\left(b_{n-1} + a_{n-1}\right).2^{2n-2} = -2^{2n-1} + \left(\overline{a_{n-1}} + \overline{b_{n-1}}\right).2^{2n-2}$$

A and B are n-bits operands, so their product is a 2n-bits number. Consequently, the most significant weight is 2n-1, and the first term -2^{2n-1} is taken into account by adding a 1 in the most significant cell of the multiplier. The implementation is shown in figure 6.25.

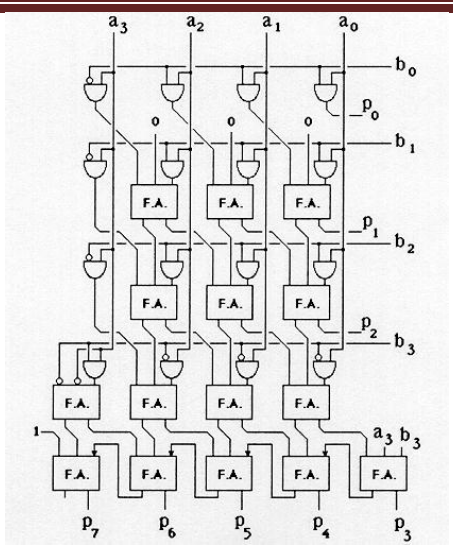


Figure-6.25: A 4-bit Baugh-Wooley Multiplier

Booth Algorithm

This algorithm is a powerful direct algorithm for signed-number multiplication. It generates a 2n-bit product and treats both positive and negative numbers uniformly. The idea is to reduce the number of additions to perform. Booth algorithm allows in the best case n/2 additions whereas modified Booth algorithm allows always n/2 additions.

Let us consider a string of k consecutive 1s in a multiplier:
..., i+k, i+k-1, i+k-2 , ..., i, i-1, ...
..., 0 , 1 , 1 , ..., 1, 0, ...

where there is k consecutive 1s.

By using the following property of binary strings:

$$2^{i+k}-2^{i}=2^{i+k-1}+2^{i+k-2}+...+2^{i+1}+2^{i}$$

the k consecutive 1s can be replaced by the following string

..., i+k+1, i+k, i+k-1, i+k-2, ..., i+1, i , i-1 , ...
..., 0 , 1 , 0 , 0 , ..., 0 , -1 , 0 , ...
k-1 consecutive 0s Addition Subtraction

In fact, the modified Booth algorithm converts a signed number from the standard 2's-complement radix into a number system where the digits are in the set {-1,0,1}. In this number system, any number may be written in several forms, so the system is called redundant.

The coding table for the modified Booth algorithm is given in Table 1. The algorithm scans strings composed of three digits. Depending on the value of the string, a certain operation will be performed.

A possible implementation of the Booth encoder is given on Figure 6.26.

Table-1: Modified Booth coding table

	BIT			M is
2^1	2^0	2^{-1}	OPERATION	multiplied
Y_{i+1}	Y_i	Y_{i-1}		by
0	0	0	add zero (no string)	+0
0	0	1	add multipleic (end of string)	+X
0	1	0	add multiplic. (a string)	+X
0	1	1	add twice the mul. (end of string)	+2X
1	0	0	sub. twice the m. (beg. of string)	-2X
1	0	1	sub. the m. (-2X and +X)	-X
1	1	0	sub . the m. (beg. of string)	-X
1	1	1	sub. zero (center of string)	-0

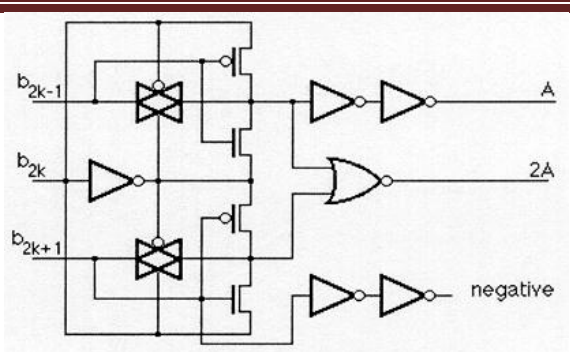


Figure-6.26: Booth encoder cell

To summarize the operation:

- Grouping multiplier bits into pairs
  - Orthogonal idea to the Booth recoding
  - Reduces the num of partial products to half
  - If Booth recoding not used have to be able to multiply by 3 (hard: shift+add)
- Applying the grouping idea to Booth Modified Booth Recoding (Encoding)
  - We already got rid of sequences of 1's no multiplication by 3
  - Just negate, shift once or twice

Wallace Trees

For this purpose, Wallace trees were introduced. The addition time grows like the logarithm of the bit number. The simplest Wallace tree is the adder cell. More generally, an n-inputs Wallace tree is an n-input operator and log2(n) outputs, such that the value of the output word is equal to the number of "1" in the input word. The input bits and the least significant bit of the output have the same weight (Figure 6.27). An important property of Wallace trees is that they may be constructed using adder cells. Furthermore, the number of adder cells needed grows like the logarithm log2(n) of the number n of input bits. Consequently, Wallace trees are useful whenever a large number of operands are to add, like in multipliers. In a Braun or Baugh-Wooley multiplier with a Ripple Carry Adder, the completion time of the multiplication is proportional to twice the number n of bits. If the collection of the partial products is made through Wallace trees, the time for getting the result in a carry save notation should be proportional to log2(n).

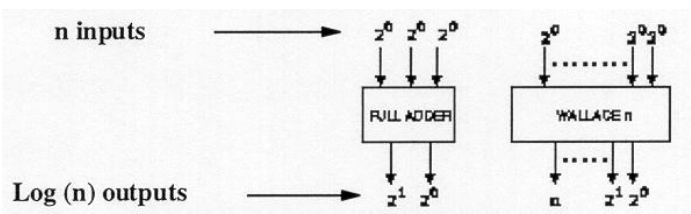


Figure-6.27: Wallace cells made of adders

Figure 6.28 represents a 7-inputs adder: for each weight, Wallace trees are used until there remain only two bits of each weight, as to add them using a classical 2-inputs adder. When taking into account the regularity of the interconnections, Wallace trees are the most irregular.

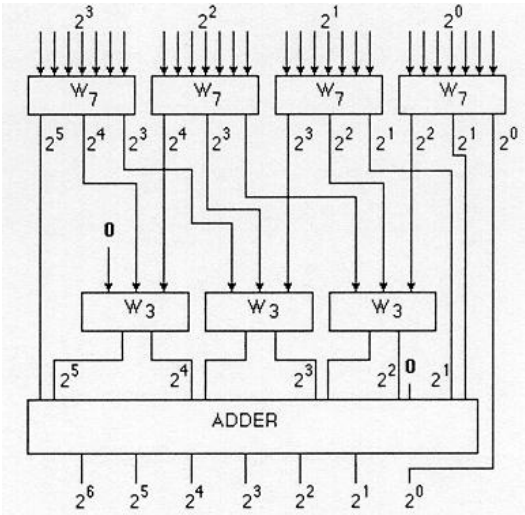


Figure-6.28: A 7-inputs Wallace tree

To summarize the operation:

The Wallace tree has three steps:

Multiply (that is - AND) each bit of one of the arguments, by each bit of the other, yielding n^2 results.
Reduce the number of partial products to two by layers of full and half adders.
Group the wires in two numbers, and add them with a conventional adder.

The second phase works as follows.

Take any three wires with the same weights and input them into a full adder.
The result will be an output wire of the same weight and an output wire with a higher weight for each three input wires.
If there are two wires of the same weight left, input them into a half adder.
If there is just one wire left, connect it to the next layer.

Memory

Objectives: At the end of this unit we will be able to understand

- System timing consideration
- Storage / Memory Elements
  - dynamic shift register
  - 1T and 3T dynamic memory
  - 4T dynamic and 6T static CMOS memory
- Array of memory cells

System timing considerations:

- Two phase non-overlapping clock
- φ_1 leads φ_2
- Bits to be stored are written to register and subsystems on φ_1
- Bits or data written are assumed to be settled before φ_2
- φ_2 signal used to refresh data
- Delays assumed to be less than the intervals between the leading edge of φ_1 & φ_2
- Bits or data may be read on the next φ_1
- There must be atleast one clocked storage element in series with every closed loop signal path

Storage / Memory Elements:

The elements that we will be studying are:

- Dynamic shift register
- 3T dynamic RAM cell
- 1T dynamic memory cell
- Pseudo static RAM / register cell
- 4T dynamic & 6T static memory cell
- JK FF circuit
- D FF circuit

Dynamic shift register:

Circuit diagram: Refer to unit 4(ch 6.5.4)

Power dissipation

- static dissipation is very small
- dynamic power is significant
- dissipation can be reduced by alternate geometry

Volatility

- data storage time is limited to 1msec or less

3T dynamic RAM cell:

Circuit diagram

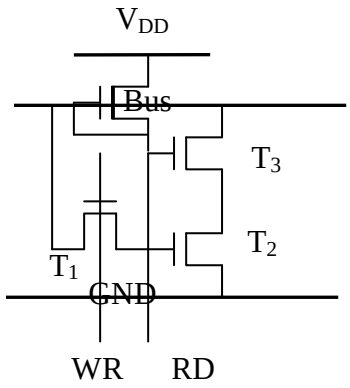


Figure 7.1: 3T Dynamic RAM Cell

Working

- RD = low, bit read from bus through T1, WR = high, logic level on bus sent to Cg of T2, WR = low again
- Bit level is stored in Cg of T2, RD=WR=low
- Stored bit is read by RD = high, bus will be pulled to ground if a 1 was stored else 0 if T2 non-conducting, bus will remain high.

Dissipation

- Static dissipation is nil
- Depends on bus pull-up & on duration of RD signal & switching frequency

Volatility

- Cell is dynamic, data will be there as long as charge remains on Cg of T2

1T dynamic memory cell:

Circuit diagram

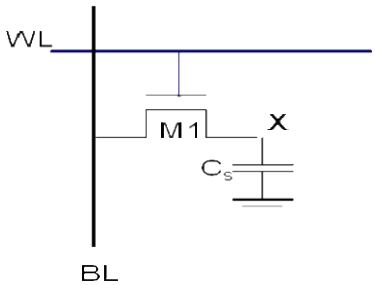


Figure 7.2: 1T Dynamic RAM Cell

Working

- Row select (RS) = high, during write from R/W line Cm is charged
- data is read from Cm by detecting the charge on Cm with RS = high
- cell arrangement is bit complex.
- solution: extend the diffusion area comprising source of pass transistor, but Cd<<< Cgchannel

- another solution : create significant capacitor using poly plate over diffusion area.
- Cm is formed as a 3-plate structure
- with all this careful design is necessary to achieve consistent readability

Dissipation

- no static power, but there must be an allowance for switching energy during read/write

- dynamic RAM need to be refreshed periodically and hence not convenient
- static RAM needs to be designed to hold data indefinitely
- One way is connect 2 inverter stages with a feedback.
- say φ_2 to refresh the data every clock cycle
- bit is written on activating the WR line which occurs with φ_1 of the clock
- bit on Cg of inverter 1 will produce complemented output at inverter 1 and true at output of inverter 2
- at every φ_2 , stored bit is refreshed through the gated feedback path
- stored bit is held till φ_2 of clock occurs at time less than the decay time of stored bit
- to read RD along with φ_1 is activated

Note:

- WR and RD must be mutually exclusive
- φ_2 is used for refreshing, hence no data to be read, if so charge sharing effect, leading to destruction of stored bit
- cells must be stackable, both side-by-side & top to bottom
- allow for other bus lines to run through the cell

4T dynamic & 6T static memory cell:

Circuit diagram

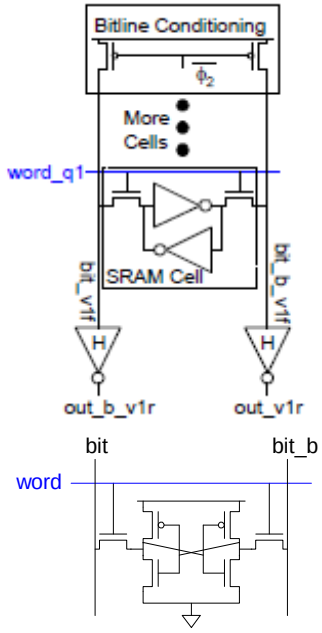


Figure 7.4: Dynamic and static memory cells

- uses 2 buses per bit to store bit and bit'
- both buses are precharged to logic 1 before read or write operation.
- write operation
- read operation

Write operation

- both bit & bit' buses are precharged to VDD with clock φ_1 via transistor T5 & T6
- column select line is activated along with φ_2
- either bit or bit' line is discharged along the I/O line when carrying a logic 0
- row & column select signals are activated at the same time => bit line states are written in via T3 & T4, stored by T1 & T2 as charge

Read operation

- bit and bit' lines are again precharged to VDD via T5 & T6 during φ_1
- if 1 has been stored, T2 ON & T1 OFF
- bit' line will be discharged to VSS via T2
- each cell of RAM array be of minimum size & hence will be the transistors
- implies incapable of sinking large charges quickly

- RAM arrays usually employ some form of sense amplifier
  - T1, T2, T3 & T4 form as flip-flop circuit
  - if sense line to be inactive, state of the bit line reflects the charge present on gate capacitance of T1 & T3
  - current flowing from VDD through an on transistor helps to maintain the state of bit lines.