UNIT I MOS TRANSISTOR PRINCIPLE

NMOS and PMOS transistors, Process parameters for MOS and CMOS, Electrical properties of CMOS circuits and device modeling, Scaling principles and fundamental limits, CMOS inverter scaling, propagation delays, Stick diagram, Layout diagrams

1.1.Introduction:

In 1958, Jack Kilby built the first integrated circuit flip-flop with two transistors at Texas Instruments. In 2008, Intel's Itanium microprocessor contained more than 2 billion transistors and a 16 Gb Flash memory contained more than 4 billion transistors. This corresponds to a compound annual growth rate of 53% over 50 years. No other technology in history has sustained such a high growth rate lasting for so long. This incredible growth has come from steady miniaturization of transistors and improvements in manufacturing processes. Most other fields of engineering involve tradeoffs between performance, power, and price. However, as transistors become smaller, they also become faster, dissipate less power, and are cheaper to manufacture.

Figure 1. Shows annual sales in the worldwide semiconductor market. Integrated circuits became a $100 billion/year business in 1994. In 2007, the industry manufactured approximately 6 quintillion transistors, or nearly a billion for every human being on the planet.

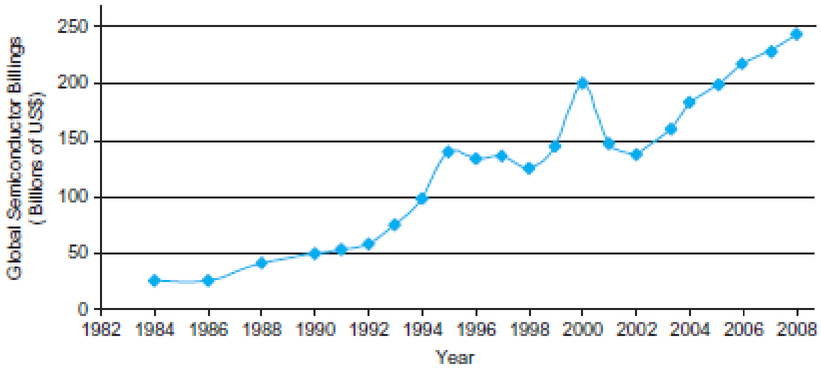


Fig.1. Size of worldwide semiconductor market

During the first half of the twentieth century, electronic circuits used large, expensive, power-hungry, and unreliable vacuum tubes. Ten years later, Jack Kilby at Texas Instruments realized the potential for miniaturization if multiple transistors could be built on one piece of silicon.

Transistors can be viewed as electrically controlled switches with a control terminal and two other terminals that are connected or disconnected depending on the voltage or current applied to the control. Soon after inventing the point contact transistor, Bell Labs developed the bipolar junction transistor. Bipolar transistors were more reliable, less noisy, and more powerefficient. Early integrated circuits primarily used bipolar transistors. Bipolar transistors require a small current into the control (base) terminal to switch much larger currents between the other two (emitter and collector) terminals. The quiescent power dissipated by these base currents, drawn even when the circuit is not switching, limits the maximum number of transistors that can be integrated onto a single die. By the 1960s, Metal Oxide Semiconductor Field Effect Transistors (MOSFETs) began to enter production. MOSFETs offer the compelling advantage that they draw almost zero control current while idle. They come in two flavors: nMOS and pMOS, using n-type and p-type silicon, respectively. With the development of the silicon planar process, MOS integrated circuits became attractive for their low cost because each transistor occupied less area and the fabrication process was simpler. Early commercial processes used only pMOS transistors and suffered from poor performance, yield, and reliability. Processes using nMOS transistors became common in the 1970s.

Figure 2.shows that the number of transistors in Intel microprocessors has doubled every 26 months since the invention of the 4004. Moore’s Law is driven primarily by scaling down the size of transistors and, to a minor extent, by building larger chips. The level of integration of chips has been classified as small-scale, medium-scale, large-scale, and very large scale. Small-scale integration (SSI) circuits, such as the 7404 inverter, have fewer than 10 gates, with roughly half a dozen transistors per gate. Medium-scale integration (MSI) circuits, such as the 74161 counter, have up to 1000 gates. Large-scale integration (LSI) circuits, such as simple 8-bit microprocessors, have up to 10,000 gates. It soon became apparent that new names would have to be created every five years if this naming trend continued and thus the term very large-scale integration (VLSI) is used to describe most integrated circuits from the 1980s onward.

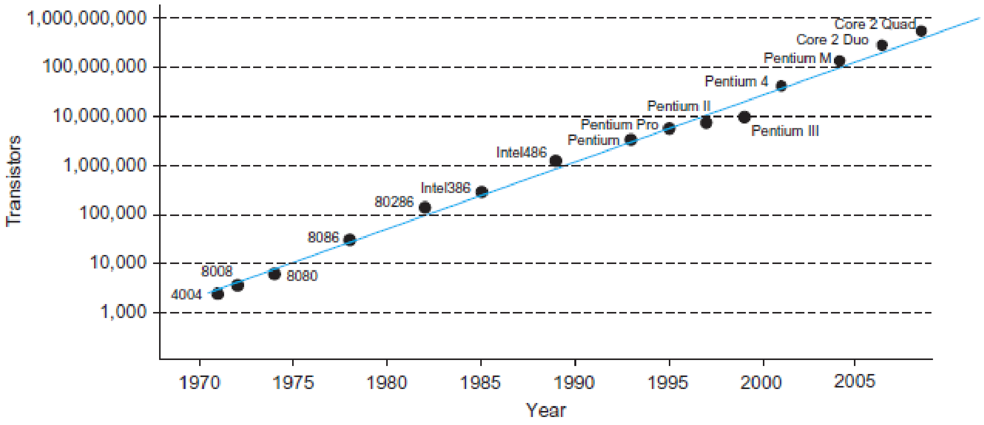


Fig.2. Transistors in Intel microprocessors

MOS Transistors:

Silicon (Si),is a semiconductor. Group IV element, so it forms covalent bonds with four adjacent atoms, as shown in Figure 3(a). Pure silicon is a poor conductor. The conductivity can be raised by introducing small amounts of impurities, called dopants, into the silicon lattice. If a Group V dopant such as arsenic with valence electrons is added, it forms an n-type semiconductor as shown in Figure 3(b). In this an excess free electron will be left in the arsenic. The free electron can carry current so that the conductivity is increased. Similarly, a Group III dopant, such as boron, has three valence electrons, as shown in Figure 3.(c).

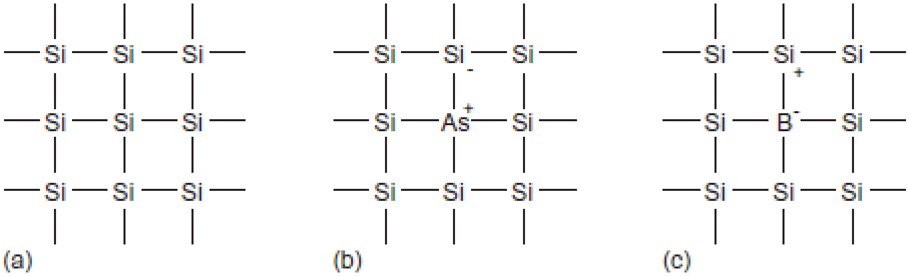


Fig.3. Silicon lattice and dopant atoms

The dopant atom can borrow an electron from a neighboring silicon atom, which in turn becomes short by one electron. That atom in turn can borrow an electron, and so forth, so the missing electron, or hole, can propagate about the lattice. The hole acts as a positive carrier so we call this a p-type semiconductor.

A junction between p-type and n-type silicon is called a diode. When the voltage on the p-type semiconductor, called the anode, is raised above the n-type cathode, the diode is forward biased and current flows. When the anode voltage is less than or equal to the cathode voltage, the diode is reverse biased and very little current flows.

A Metal-Oxide-Semiconductor (MOS) structure is created by superimposing several layers of conducting and insulating materials to form a sandwich-like structure. These structures are manufactured using a series of chemical processing steps involving oxidation of the silicon, selective introduction of dopants, and deposition and etching of metal wires and contacts. Transistors are built on nearly flawless single crystals of silicon, which are available as thin flat circular wafers of 15–30 cm in diameter. CMOS technology provides two types of transistors (also called devices): an n-type transistor (nMOS) and a p-type transistor (pMOS). Transistor operation is controlled by electric fields so the devices are also called Metal Oxide Semiconductor Field Effect Transistors (MOSFETs) or simply FETs. Cross-sections and symbols of these transistors are shown in Figure.4.The n+ and p+ regions indicate heavily doped n- or p-type silicon.

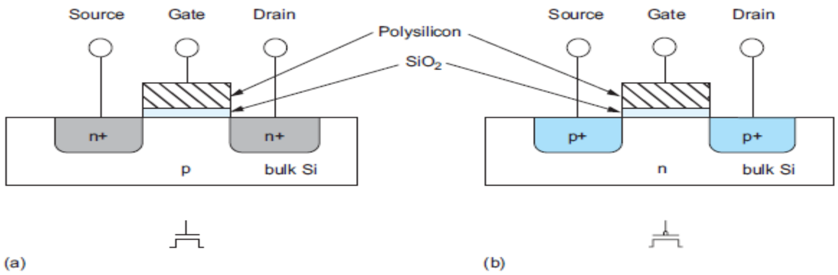


Fig.4. nMOS transistor (a) and pMOS transistor (b)

Each transistor consists of a stack of the conducting gate, an insulating layer of silicon dioxide (SiO2, better known as glass), and the silicon wafer, also called the substrate, body, or bulk. Gates of early transistors were built from metal, so the stack was called metal oxide semiconductor, or MOS. Since the 1970s, the gate has been formed from polycrystalline silicon (polysilicon), but the name stuck. An nMOS transistor is built with a p-type body and has regions of n-type semiconductor adjacent to the gate called the source and drain. They are physically equivalent and for now we will regard them as interchangeable. The body is typically grounded. A pMOS transistor is just the opposite, consisting of p-type source and drain regions with an n-type body.

The gate is a control input: It affects the flow of electrical current between the source and drain. Consider an nMOS transistor, When the gate of an nMOS transistor is 1, the transistor is ON and there is a conducting path from source to drain. When the gate is low, the nMOS transistor is OFF and almost zero current flows from source to drain. For a pMOS transistor, the situation is again reversed. A pMOS transistor is just the opposite, being ON when the gate is low and OFF when the gate is high. Notice that the symbol for the pMOS transistor has a bubble on the gate, indicating that the transistor behavior is the opposite of the nMOS.The positive voltage is usually called V_{DD} or POWER and represents a logic 1 value in digital circuits. The low voltage is called GROUND (GND) or V_{SS} and represents a logic 0,as shown in fig.5.

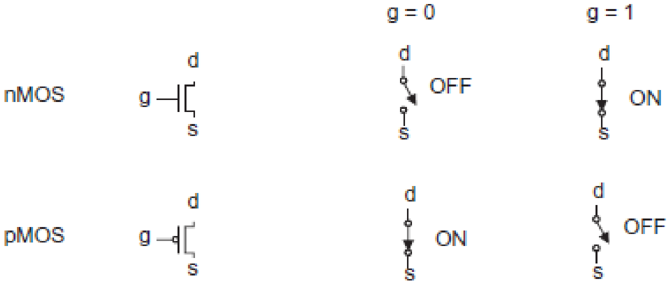


Fig.5. Transistor symbols and switch-level models

1.1.1 Modes Of Operation Of MOS transistor:

In Figure 6.(a) , a negative voltage is applied to the gate, so there is negative charge on the gate. The mobile positively charged holes are attracted to the region beneath the gate. This is called the accumulation mode.

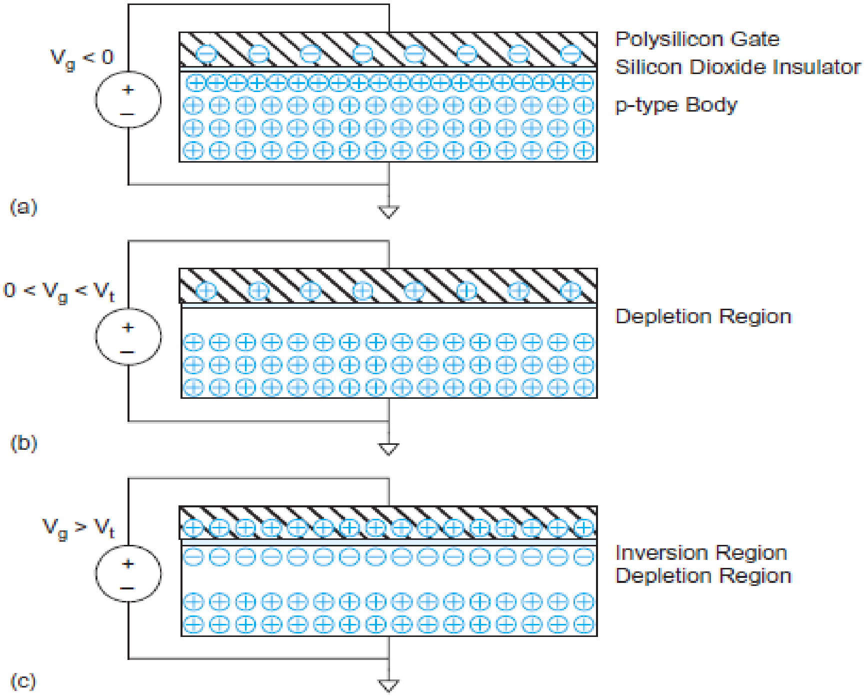


Fig.6. MOS structure demonstrating (a) accumulation, (b) depletion, and (c) inversion

In Figure 6. (b), a small positive voltage is applied to the gate, resulting in some positive charge on the gate. The holes in the body are repelled from the region directly beneath the gate, resulting in a depletion region forming below the gate. In Figure 6.(c), a higher positive potential exceeding a critical threshold voltage V_t is applied, attracting more positive charge to the gate. The holes are repelled further and some free electrons in the body are attracted to the region beneath the gate. This conductive layer of electrons in the p-type body is called the inversion layer.

1.1.2.Behavior of nMOS with different voltages:

In Figure 7.(a), the gate-to-source voltage V_{gs} is less than the threshold voltage. The source and drain have free electrons. The body has free holes but no free electrons. Suppose the source is grounded.

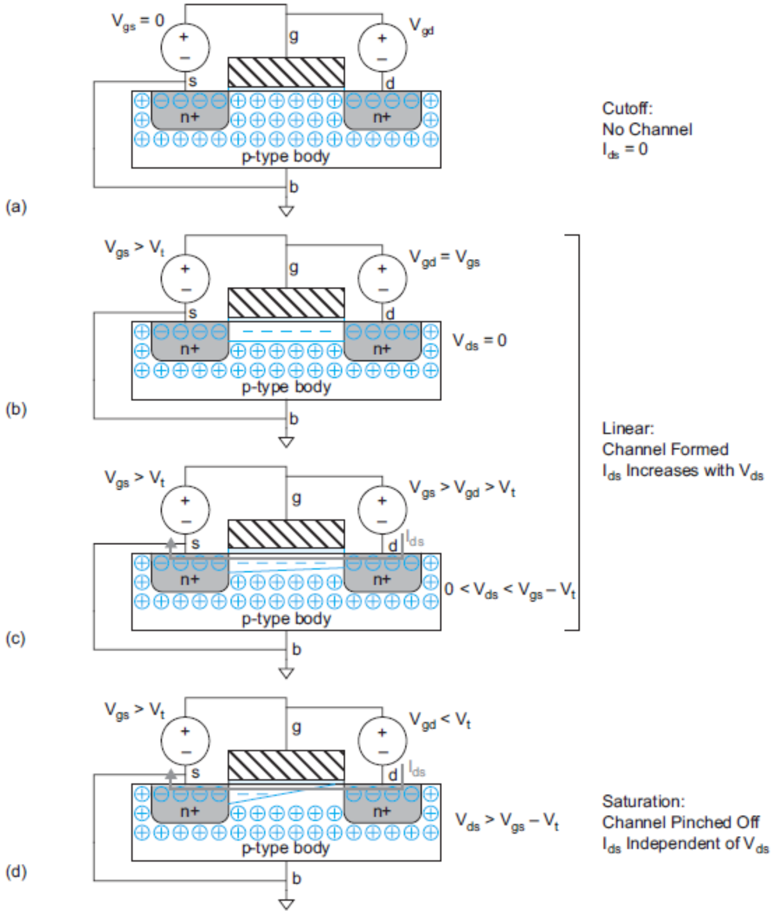


Fig.7 nMOS transistor demonstrating cutoff, linear, and saturation regions of operation

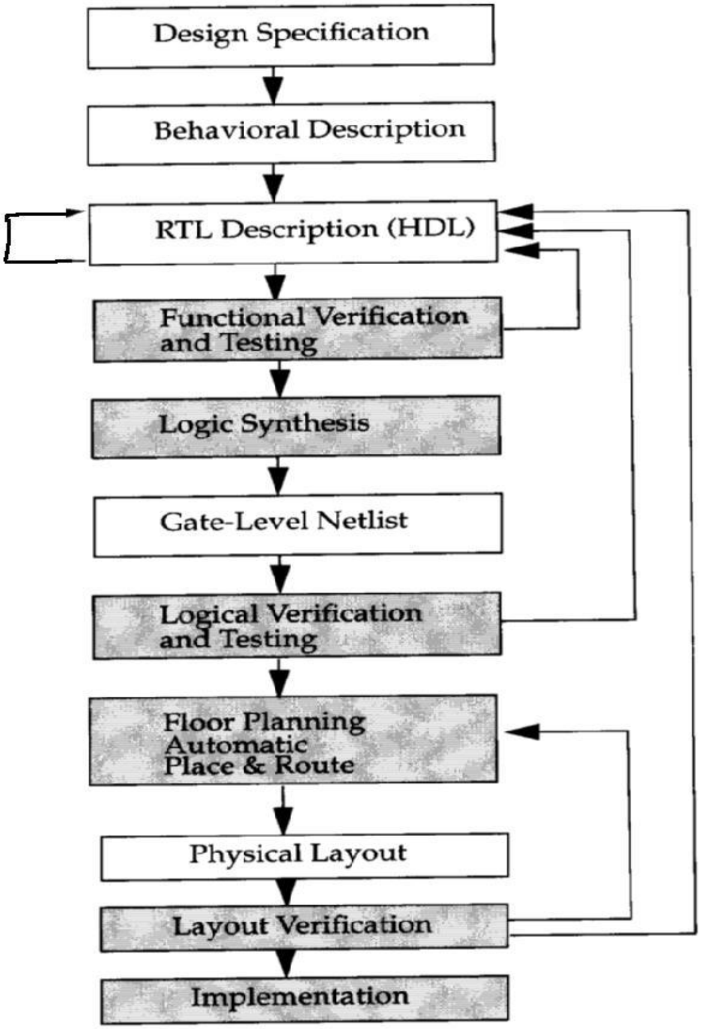
The junctions between the body and the source or drain are zero-biased or reverse-biased, so little or no current flows. We say the transistor is OFF, and this mode of operation is called cutoff. In Figure 7.(b), the gate voltage is greater than the threshold voltage. Now an inversion region of electrons (majority carriers) called the channel connects the source and drain, creating a conductive path and turning the transistor ON. The number of carriers and the conductivity increases with the gate voltage. The potential difference between drain and source is $V_{ds} = V_{gs} - V_{gd}$. If $V_{ds} = 0$ (i.e., $V_{gs} = V_{gd}$),there is no electric field tending to push current from drain to source.

When a small positive potential V_{ds} is applied to the drain (Figure 7. (c)), current I_{ds} flows through the channel from drain to source. This mode of operation is termed linear, resistive, triode, non saturated, or unsaturated; the current increases with both the drain voltage and gate voltage. If V_{ds} becomes sufficiently large that $V_{gd} < V_t$, the channel is no longer inverted near the drain and becomes pinched off (Figure 7.(d)). However, conduction is still brought about by the drift of electrons under the influence of the positive drain voltage. As electrons reach the end of the channel, they are injected into the depletion region near the drain and accelerated toward the drain. Above this drain voltage the current Ids is controlled only by the gate voltage and ceases to be influenced by the drain. This mode is called saturation.

1.1.3.VLSI Design flow

A typical design flow for designing VLSI IC circuits is shown in Figure. Unshaded blocks show the level of design representation; shaded blocks show processes in the design flow.

- In any design specification are written first ,specifications describe abstractly the functionality,interface and overall architecture of the digital circuit to be designed.
- Behavioral description is used to analyse design in term of functions,performance compliance to standards and other high level issues.
- Behaviours description is manually converted to RTL(Register transfer level) description in HDL.W e have to describe data flow that will implement the desired digital circuit,then design process is done.
- Logic synthesis tools consists RTL Description to gate level netlist (input to automatic place and route tool which creates layout)
- Gate level netlist is a description of the circuit in terms of gates and connection between them
- Logic synthesis tools ensure gate level netlist meets timing,area,power specification
- Implementation helps the design to convert behavioural description to a final Ic chip.



1.2.Process parameters for MOS and CMOS (or)Layout Design Rule:

Layout rules are also referred to as design rules or ground rules.

- It is considered as prescription for preparation of photo masks in the fabrication process.
- The main objective of layout rules is to build reliably
- Functional circuits in a small area.
- It provides a communication link between circuit designer and process engineer during the manufacturing phase.
- The rules are defined in terms of feature sizes(width) separations and overlaps.

Layer Representation:

CMOS process use the following features

- Two different substrate
- Doped region of both P and N material
- Transistor gate electrode
- Inter layer contacts.

These layers for CMOS process are represented in various figures in terms of

- Colour scheme
- Varying stipple pattern
- Varying line styles

Example: N-well---brown

Thin oxide---green

Poly silicon---red

Contact cut---black

N-well rules:

- N-well rule is based on the MOSIS CMOS scalable rules
- MOSIS is expressed in terms of λ.
- Industry actually uses the μ(micron) design rules and codes designs in terms of these dimensions.
- MOSIS rules allow some degree of scaling between process.

Example:

N-well	λ rule	λ/ μ RULE	μ rule
Min size	10 λ	5 μ	2 μ
Min spacing(wells at same potential)	6 λ	3 μ	2 μ
Min spacing(wells at different potential)	8 λ	4 μ	2 μ

Design rules background:

1.Well rules:

- The well is usually a deeper than the transistor source/drain.
- Hence it is necessary for the outside dimension to provide sufficient clearance between the N-well edges and the adjacent diffusions.
- The inside clearance is determined by the transition of field oxide across the well boundary.
- Some process may permit zero inside clearance this leads to 'birds-beaks'(transition from the thick film to thin film)
- N-well sheet resistance can be several kΩs per square. This prevent the excess voltage drop due to the substrate current.

2.Transistor rules:

The source and drain diffusion is masked by the poly region. It is essential for the poly to completely cross active otherwise the transistor will be shorted by a diffused path between source and drain.

To ensure this condition is satisfied poly is required to extend beyond the edges of the diffusion region. This is called as gate extension. Figure 34.(a) shows the mask construction for the final structures that appear in Figure 34.(b).

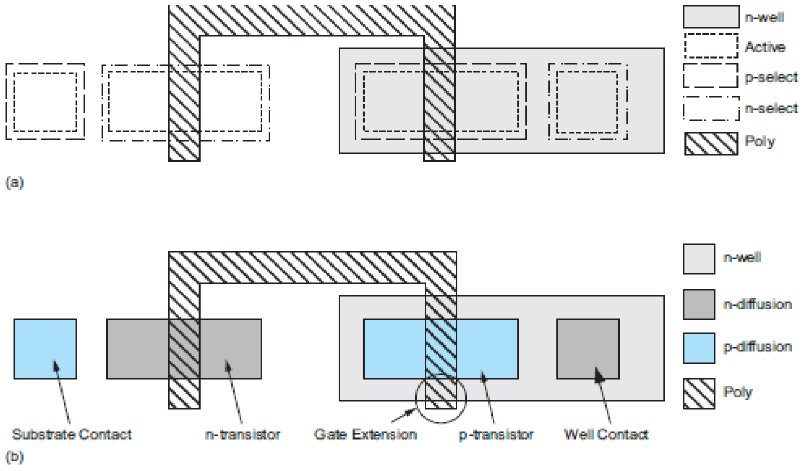


Fig.34.CMOS n-well process transistor and well/substrate contact construction

Contact rules:

The generally available contacts are

- Metal to p-active
- Metal to n-active
- Metal to poly silicon
- Metal to well (or) substrate.

The substrate is divided into well region and each isolated well must be tied to the approximate supply voltage i.e., n-well tied to VDD and p-well tied to GND .

- A metal makes poor connection to be lightly doped substrate (or) well hence a heavily doped active region is placed between the contact.
- The spit or merged contact is equivalent to 2 separate metal diffusion contacts that are strapped together with metal.
- This structure is used to tie transistor source to either the substrate or the well.

Guard Rings:

- They are p+ diffusion in the p substrate and n+ diffusion in the N-well.
- It is used to collect the injected minority carriers.
- n+ guard rings must be tied to VDD and p+ guard rings to Vss.

Metal rules:

- Metal spacing may vary with the width of the metal line.
- At some width the metal spacing may be increased. This is due to etch characteristics of small versus large metal wires.
- The rules are applied to closely spaced parallel metal lines and maximum metal width rules are also applied.

Via rules:

- Process may vary in whether they allow vias to be placed over polysilicon and diffusion regions
- Some processes allow via to be placed within these areas, but donot allow vias to be strddle the boundary of polysilicon or diffusion

Some additional rules are

- Extension of polysilicon in the direction that metal wires exit a contact.
- Differing p and n transistor gate length.

- Differing gate poly extension.

Passivation (or) overglass:

This is a protective glass layer that covers the final chip

Scribe line:

The scribe line is a specifically designed structure that surrounds the completed chip and in the point at which the chip is cut with a diamond saw. It is designed to prevent the ingress of contaminant from side of the chip.

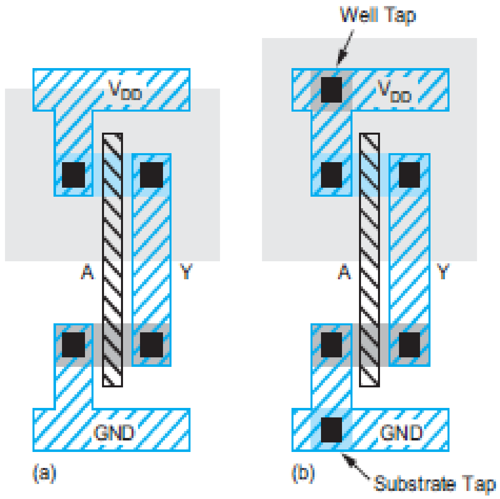


FIGURE 1.41 Inverter cell layout

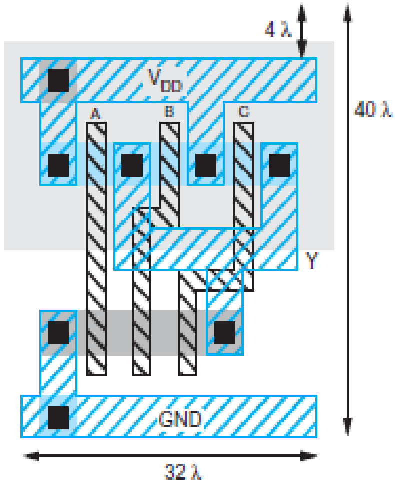


FIGURE 1.42 3-input NAND standard cell gate layouts

1.3.Electrical properties of CMOS Circuits and Device Modeling:

1.3.1.Ideal I-V characteristics (or) MOS device design equations (or) basic DC equation:

MOS transistors have three regions of operation:

1. Cutoff or sub threshold region
2. Linear region
3. Saturation region

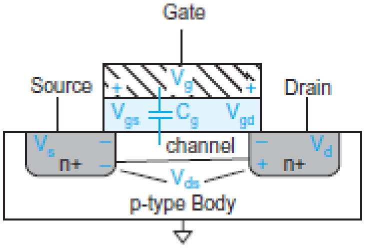
Let us derive a model relating the current and voltage (I-V) for an nMOS transistor in each of these regions. The model assumes that the channel length is long enough that the lateral electric field (the field between source and drain) is relatively low, which is no longer the case in nanometer devices. This model is variously known as the long-channel, ideal, first-order, or

Shockley model. Subsequent sections will refine the model to reflect high fields, leakage, and other nonidealities.

The long-channel model assumes that the current through an OFF transistor is 0.When a transistor turns ON ($V_{gs} > V_t$), the gate attracts carriers (electrons) to form a channel. The electrons drift from source to drain at a rate proportional to the electric field between these regions. Thus, we can compute currents if we know the amount of charge in the channel and the rate at which it moves. We know that the charge on each plate of a capacitor is Q =CV. Thus, the charge in the channel $Q_{channel}$ is,

$$Q_{channel} = C_g(V_{gc} - V_t)$$

where C_g is the capacitance of the gate to the channel and $V_{gc} = V_t$ is the amount of voltage attracting charge to the channel beyond the minimum required to invert from p to n. The gate voltage is referenced to the channel, which is not grounded. If the source is at V_s and the drain is at V_d , the average is $V_c=(V_s /V_d)/2 = V_s + V_{ds} /2$. Therefore, the mean difference between the gate and channel potentials V_{gc} is $V_g - V_c = V_{gs} - V_{ds} /2$, as shown in Figure 8.



Average gate to channel potential:

$V_{gc} = (V_{gs} + V_{gd})/2 = V_{gs} - V_{ds}/2$

Fig.8 Average gate to channel voltage

We can model the gate as a parallel plate capacitor with capacitance proportional to area over thickness. If the gate has length L and width W and the oxide thickness is t_{ox}, as shown in Figure 9. the capacitance is

$$C_g = k_{ox}\varepsilon_0 \frac{WL}{t_{ox}} = \varepsilon_{ox}\frac{WL}{t_{ox}} = C_{ox}WL$$

where ε0 is the permittivity of free space, 8.85×10^{-14} F/cm, and the permittivity of SiO2 is k_{ox} =3.9 times as great. Often, the ε_0 / t_{ox} term is called C_{ox}, the capacitance per unit area of the gate oxide.

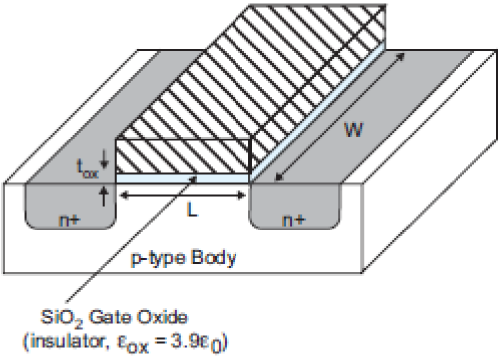


Fig.9 Transistor dimensions

Some nanometer processes use a different gate dielectric with a higher dielectric constant. In these processes, we call t_{ox} the equivalent oxide thickness (EOT), the thickness of a layer of SiO2 that has the same C_{ox}. In this case, t_{ox} is thinner than the actual dielectric. Each carrier in the channel is accelerated to an average velocity, v, proportional to the lateral electric field, i.e., the field between source and drain. The constant of proportionality μ is called the mobility.

$$v = \mu E$$

The electric field E is the voltage difference between drain and source Vds divided by the channel length

$$E = \frac{V_{ds}}{L}$$

The time required for carriers to cross the channel is the channel length divided by the carrier velocity: L/v. Therefore, the current between source and drain is the total amount of charge in the channel divided by the time required to cross

$$I_{ds} = \frac{Q_{channel}}{L/v}$$

The cutoff region:

$I_{ds}=0, \qquad V_{gs} \leq V_t$

The nonsaturation or linear or triode region:

I_{ds} varies linearly with V_{gs} and V_{ds} when the quadratic term $V^2_{ds}/2$ is very small.

$I_{ds}=\beta[(V_{gs}-V_t)V_{ds}-V^2_{ds}], \qquad 0<V_{ds}<V_{gs}-V_t$

The saturation region:

$$I_{ds}= \beta(V_{gs}-V_t)^2/2, \qquad 0<V_{gs}-Vt<V_{ds}$$

Where Ids is the drain to source current, V_{gs} is the gate to source voltage, V_t is the device threshold and β is the MOS transistor gain factor.

$$\beta=\frac{\mu\varepsilon_{ox}}{tox}(w/L)$$

Where μ is the effective mobility of the carriers in the channel, ε is the permittivity of the gate insulator, t_{ox} is the thickness of the gate insulator, W is the width of the channel and L is the length of the channel.

Where, $C_{ox}= \varepsilon_{ox}/t_{ox}$

Figure 10.(a) shows the I-V characteristics for the N and P transistors. According to the first-order model, the current is zero for gate voltages below V_t. For higher gate voltages, current increases linearly with V_{ds} for small V_{ds} .

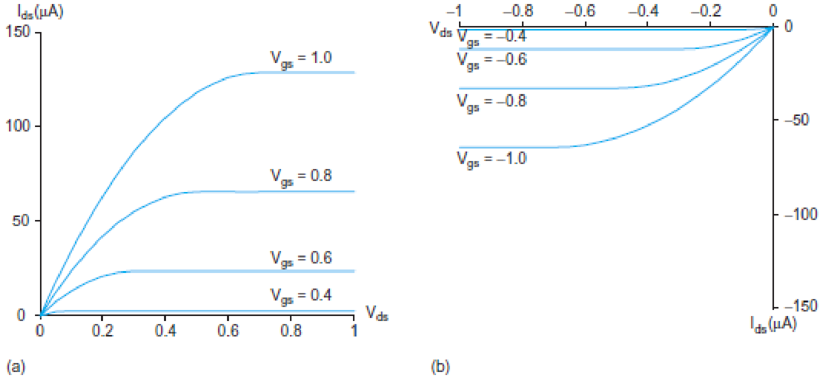


Fig.10_ I-V characteristics (a) nMOS and (b) pMOS transistors

1.3.2.Non ideal IV effects:second order effects:

The saturation current increases less than quadratically with increasing V_{gs}. This is caused by two effects: velocity saturation and mobility degradation. At high lateral field strength (V_{ds}/L), carriers velocity ceases to increase linearly with field strength. This is called velocity saturation and results in lower Ids than expected at high V_{ds}. At high vertical field strength the carriers scatter more often. This mobility degradation effect also leads to less current than expected at high V_{gs}. The saturation current of the non ideal transistor increases slightly with V_{ds}. This is caused by channel length modulation, in which higher V_{ds} increases the size of the depletion region around the drain and thus effectively shortens the channel.

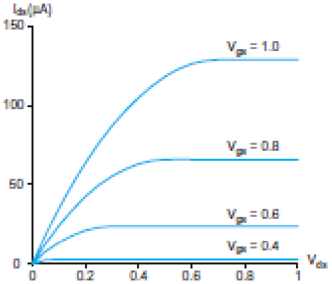


Fig.11 ideal I-V characteristics

In the above fig 11.V_{gs}<V_t, the current drops off exponentially rather than abruptly becoming zero. This is called sub threshold conduction. The threshold voltage itself is influenced by the voltage difference between the source and body this is called the body effect. The source and drain diffusions are reverse –biased diodes and also experience junction leakage into the substrate or well .The current into the gate I_g is ideally ‘0’.However as the thickness of gate oxides reduces to only a small number of atomic layers, electrons tunnel through the gate causing some gate current.

Both mobility and threshold voltage decrease with rising temperature. The mobility effect is most important for ON transistors, resulting in lower Ids at high temperature. The threshold effect is most important for OFF transistors, resulting in higher leakage current at high temperature. Clearly MOS characteristics degrade with temperature.

Velocity saturation and mobility degradation:

Figure 12. Shows that carrier drift velocity and current increase linearly with the lateral electric field E_{lat}=V_{ds}/L between source and drain. At high field strength, drift velocity rolls due to carrier scattering and eventually saturates at V_{sat}.

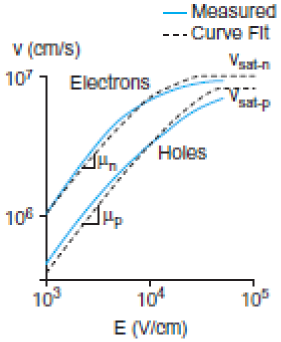


Fig.12 Carrier velocity vs. electric field

$$V=\frac{\square E_{lat}}{1+\frac{E_{lat}}{E_{sat}}}$$

The saturation current is $I_{ds}= \mu C_{ox}\frac{w}{L}\frac{(Vgs-Vt)^2}{2}$

If the transistor were completely velocity saturated $=V_{sat}$

$I_{ds}=C_{ox}W(V_{gs}-V_t)V_{sat}$

Overall the model is based on three parameters.,

$$I_{ds}=\begin{cases} 0, V_{gs} < V_t & \text{cutoff} \\ I_{dsat}\frac{V_{ds}}{V_{dsat}}, V_{ds} < V_{sat} & \text{linear} \\ I_{dsat}, V_{ds} > V_{dsat} & \text{saturation} \end{cases} \qquad \text{Where, } I_{dast}=P_c\frac{\beta}{2}(V_{gs}-{}^{V_t})^{\alpha}$$

$$V_{dsat}=p_v(V_{gs}-{}^{V_t})^{\alpha/2}$$

As channel lengths become shorter, the lateral field increases and transistors become more velocity saturated. If the supply voltage is held constant.

The low field mobility of holes is much lower than that of electrons, so pMOS experience less velocity saturation than nMOS for a given V_{DD}.

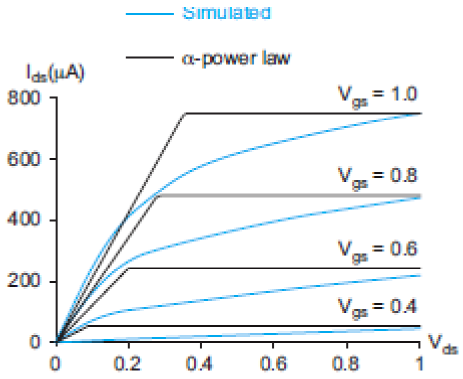


Fig.13.Comparison of α power law model with simulated transistor behavior

Strong vertical electric fields resulting from large V_{gs} cause the carriers to scatter against the surface and also reduce the carrier mobility µ.This effect is called mobility degradation.

Channel length modulation:

The reverse biased p-n junction between the drain and body forms a depletion region effectively shortens the channel length to $L_{eff}=L-L_d$. To avoid introducing the body voltage into our calculations assume the source voltage is close to the body voltage. So V_{db}~V_{ds} , hence increasing V_{ds} decreases the effective channel length. Shorter channel length results in higher current. Thus Ids increases with V_{ds} in saturation as shown below.

$$I_{ds}=\frac{\beta}{2}(V_{gs}-{}^{V_t})^2(1+\lambda V_{ds})$$

Body effect:

The threshold voltage V_t is not constant with respect to the voltage difference between the substrate and the source of the MOS transistor. This is known as the substrate bias effect or body effect.

$$V_t=V_{fb}+2\varphi_b\frac{\sqrt{2\varepsilon s_i qNA\ (2\varphi b+|Vsb\ |}}{Cox}$$

$$V_t=V_{t0}+\gamma\ [\sqrt{(2\varphi b+|Vsb|)}-\sqrt{2\varphi b}]$$

$$\gamma=\frac{tox}{\varepsilon ox}\sqrt{2q\varepsilon s_{iNA}}=\frac{1}{Cox}\sqrt{2q\varepsilon s_i NA}$$

subthreshold conduction:

The ideal transistor I-V model assumes current only flows from source to drain when $V_{gs}>V_t$. In real transistor current doesnot abruptly cutoff below threshold. but rather drops off exponentially as given in equation.

$$I_{ds}=I_{ds0}\ e^{\frac{Vgs\ -V}{nV_T}}(1-e^{\frac{-Vds}{V_T}})$$

$$I_{ds0}=\beta v_T{}^2 e^{1.8}$$

This condition is also known as leakage and often results in undesired current when a transistor is normally OFF. Here leakage is 0 if V_{ds}=0, subthreshold conduction is used to advantage in very low-power analog circuits.

Junction leakage:

The p-n junction between diffusion and substrate or well form diodes as shown in fig. The well to substrate junction is another diode . The substrate and well are tied to GND or V_{DD} to ensure these diodes remain reverse-biased. However reverse biased diodes still conduct a small amount of current I_D.

$$I_D=Is(\frac{V_D}{e^{V_T-1}})$$

Where I_s depends on doping levels and on the area and perimeter of the diffusion region and V_d is the diode voltage.

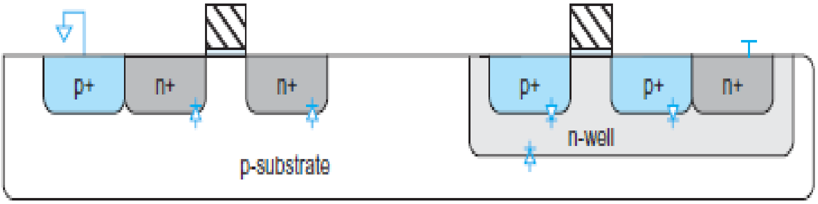


Fig.14 .Substrate to diffusion diodes in CMOS circuits

When a junction is reverse biased by significantly more than the thermal voltage the leakage is just –Is generally in the 0.1 to 0.01$\frac{\text{fA}}{\mu m^2}$ range.

Tunneling:

When the gate oxide is very thin, a current can flow from gate to source or drain by electron tunneling through the gate oxide. This current is proportional to the area of the gate of the transistor.

$I=C_1WLE_{ox}{}^2e^{\frac{-E_0}{E_{OX}}}$, E_O and C_1 are constant.

Temperature Dependence:

Transistor characteristics are influenced by temperature carrier mobility decreases with temperature.

$$\mu(T)=\mu(T_r)(\frac{T}{T_r})^{-k\,\mu}$$

where T= absolute temperature

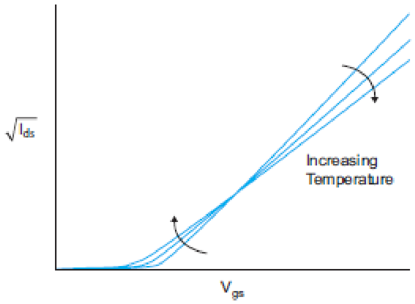
Tr=room temperature

k μ=fitting parameter generally in the range of 1.2 to 2.0.

The magnitude of the threshold voltage decreases linearly with temperature and may be approximated by,

$$V_t(T)=V_t(T_r)-K_{vt}(T-T_r)$$

K_{vt} is typically in the range of 0.5 to 3.0mV/k. Junction leakage also increases with temperatures because I_s is strongly temperature dependent.



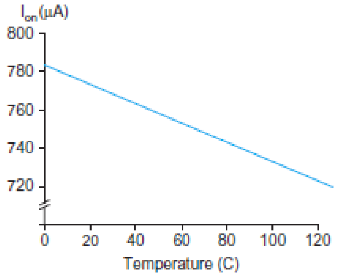


Fig.15. I–V characteristics of nMOS transistor in saturation at various temperatures

I_{dsat} vs temperature

From fig where ON current decreases and OFF current increases with temperature, second figure shows how the ON current Idsat decreases with temperature. Therefore circuit performance is generally worst at high temperature. This is called a negative temperature coefficient. Circuit performance can be improved by cooling.

Geometry Dependence:

The layout designer drawn transistors with width and length w_{drawn} and L_{drawn}. The actual gate dimensions may differ by some factors X_W and X_L. The effective transistor lengths and widths are

$$L_{eff}=L_{drawn}+X_L-2L_D$$

$$W_{eff}=W_{drawn}+X_W-2W_D$$

1.3.3.C-V characteristics or capacitance Estimation:

Each terminal of an MOS transistor has capacitance to the other terminals. In general, these capacitances are non linear and voltage dependent(C-V).

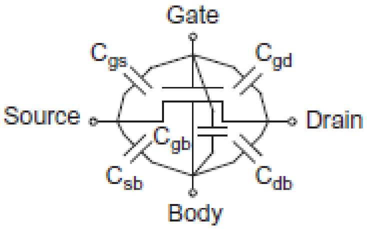


Fig.16 Capacitance of an MOS transistor

An MOS transistor can be viewed as a four-terminal device with capacitances between each terminal pair, as shown in Figure 16. The gate capacitance includes an intrinsic component (to the body, source and drain, or source alone, depending on operating regime) and overlap terms with the source and drain. The source and drain have parasitic diffusion capacitance to the body.

Simple MOS capacitance models:

The gate of an MOS transistor is a capacitor. Indeed its capacitance is necessary to attract charge to invert the channel, So high gate capacitance is required to obtain high I_{ds}. Therefore the capacitance is

$$C_g=C_{ox}wL$$

A capacitor is a two terminal device. When the transistor is ON the channel extends from the source. Most transistor used in logic are of minimum length because this result in greatest speed and lowest power dissipation. Thus taking this minimum L as a constant for a particular process we can define

$$C_g=C_{permicron}.W$$

Where, $C_{permicron}=C_{ox}L=\frac{\epsilon ox}{tox}L$

In addition to the gate, the source and drain also have capacitances. These capacitance are not fundamental to operation of the devices .but do impact circuit performance and hence are called parasitic capacitors. They arise from the reverse biased p-n junction between the source or drain diffusion and the body and hence are also called diffusion capacitance C_{sb} and C_{db}.

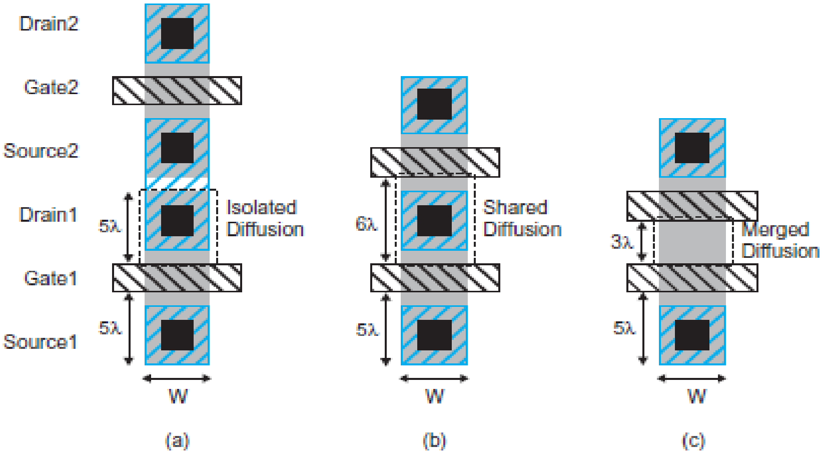


Fig.17. Diffusion region geometries

The size of these junctions depends on the area and perimeter of the source and drain diffusion, the depth of the diffusion, the doping levels, and the voltage. As diffusion has both high capacitance and high resistance, it is generally made as small as possible in the layout. There are three types of diffusion regions frequently seen illustrated with the two series transistors in fig.17. The average capacitance of each of these types of regions can be calculated or measured from simulation as a transistor switches between VDD and GND.

Detailed MOS gate capacitance model:

The gate capacitance has two components: the intrinsic capacitance and the overlap capacitance.

$$C_o=WLC_{ox}$$

1.cutoff:

When the transistor is OFF ,the channel is not inverted and charge on the gate is matched with opposite charge from the body. This is called C_{gb}, the gate to body capacitance. As V_{gs} increases but remains below a threshold, a depletion region forms at the surface. This effectively moves the bottom plate downward from the oxide, reducing the capacitance.

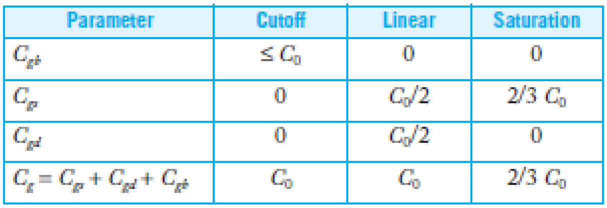
2.Linear:

When $V_{gs}>V_t$,the channel inverts and again serves as a good conductive bottom plate. However the channel is connected to the source and drain, rather than the body. At low values of V_{ds} the channel charge is roughly shared between source and drain. So a greater fraction of the capacitance is attributed to the source and drain becomes less inverted, so a greater fraction of the capacitance is attributed to the source and a smaller fraction to the drain.

3.Saturation:

At $V_{ds}>V_{gs}-V_t$, the transistor saturates and the channel pinches off. At this point all the intrinsic capacitance is to the source. Because of pinch off, the capacitance in saturation reduces to $C_{gs}=\frac{2}{3}c_0$ for an ideal transistor.The behavior in these three regions can be approximated as shown in Table .1.

Table:1.Approximation for intrinsic MOS gate capacitance

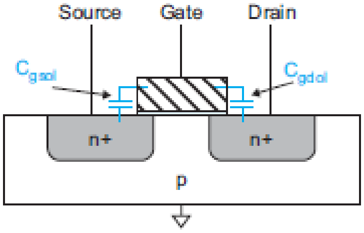


Parameter	Cutoff	Linear	Saturation
C_{gb}	$\leq C_0$	0	0
C_{gs}	0	$C_0/2$	$2/3\ C_0$
C_{gd}	0	$C_0/2$	0
$C_g = C_{gs} + C_{gd} + C_{gb}$	C_0	C_0	$2/3\ C_0$

The gate overlaps the source and drains by a small amount in a real and also has fringing fields terminating on the source and drain. This leads to additional overlap capacitances as shown in fig.18.

$$C_{gsol\,(overlap\,)} = C_{gsol}\,W$$

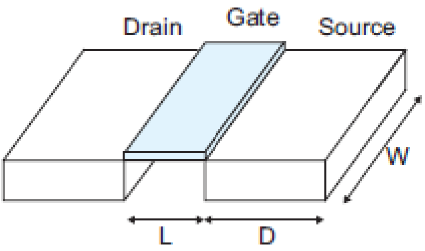
$$C_{gdol\,(overlap\,)} = C_{gdol}\,W$$



Fig,18 Overlap capacitance

Detailed MOS diffusion capacitance model:

The reverse biased p-n junction between the source diffusion and the body contributes parasitic capacitance. The capacitance depends on both the area ‘As’ and side wall perimeter ‘ps’ of the source diffusion region. The area is AS=W.D. The perimeter is PS=2.W+2D of this perimeter, as shown in fig.



The total source parasitic capacitance is C_{sb}=AS.C_{jbs}+PS.C_{jbssw}

Where C_{jb},has units of capacitances/area and C_{jbssw} has units of capacitance/length.

The area junction capacitance is $C_{jbs}=C_J(1+\frac{vSB}{\psi 0})^{-M_J}$

C_J is the junction capacitance at zero bias and is highly process dependent. M_J is the junction grading coefficient, typically in the range of 0.5 to 0.33 depending on the abruptness of the diffusion junction. Ψ_0 is the built-in potential that depends on doping levels.

Where $\psi 0$ =$v_T$ln$\frac{N_A N_D}{{n_i}^2}$

V_T is the thermal voltage

K=1.380*10^{-23}J/K

$q=1.602*10^{-19}c$

The sidewall capacitance term is of a similar form but uses different coefficients.

$$C_{jbssw}=C_{Jsw}(1+\frac{vSB}{\psi 0})^{-M}{}_{Jsw}$$

Diffusion region were historically used for short wires called runners in process with limited numbers of metal levels. Diffusion capacitance and resistance are large enough that such practice is now discouraged.

1.3.4.The complementary CMOS inverter-DC transfer characteristics:

A complementary CMOS inverter consists of a p-type and an n-type device connected in series. The DC transfer characteristics of the inverter are a function of the output voltage (V_{out}) with respect to the input voltage (V_{in}).Shown in fig.19

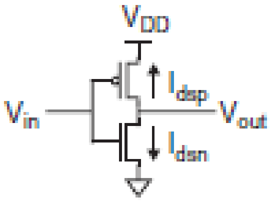


Fig.19. A CMOS inverter

Table.2.Relationships between voltages for the three regions of operation of a CMOS inverter

	Cutoff	Linear	Saturated
nMOS	$V_{gsn} < V_{tn}$	$V_{gsn} > V_{tn}$	$V_{gsn} > V_{tn}$
	$V_{in} < V_{tn}$	$V_{in} > V_{tn}$	$V_{in} > V_{tn}$
		$V_{dsn} < V_{gsn} - V_{tn}$	$V_{dsn} > V_{gsn} - V_{tn}$
		$V_{out} < V_{in} - V_{tn}$	$V_{out} > V_{in} - V_{tn}$
pMOS	$V_{gsp} > V_{tp}$	$V_{gsp} < V_{tp}$	$V_{gsp} < V_{tp}$
	$V_{in} > V_{tp} + V_{DD}$	$V_{in} < V_{tp} + V_{DD}$	$V_{in} < V_{tp} + V_{DD}$
		$V_{dsp} > V_{gsp} - V_{tp}$	$V_{dsp} < V_{gsp} - V_{tp}$
		$V_{out} > V_{in} - V_{tp}$	$V_{out} < V_{in} - V_{tp}$

In this table.2 V_{tn} is the threshold voltage of the n-channel, and V_{tp} is the threshold voltage of the p-channel device. Note that V_{tp} is negative. The equations are given both in terms of V_{gs}/V_{ds} and V_{in}/V_{out}. As the source of the nMOS transistor is grounded, $V_{gsn}=V_{in}$ and $V_{dsn}=V_{out}$. As the source of the pMOS transistor is tied to V_{DD}, $V_{gsp}=V_{in}-V_{DD}$ and $V_{dsp}=V_{out}-V_{DD}$.

The objective is to find the variation in output voltage as a function of a function of the input voltage. The pMOS transistor is 2-3 times as wide as the nMOS transistor so βn= βp. The given fig.20 shows I_{dsn} and I_{dsp} in terms of V_{dsn} and V_{dsp} for various values of V_{gsn} and V_{gsp}.Figure 20(c) shows the inverter DC transfer characteristics. The supply current $I_{DD}=I_{dsn}=|I_{dsp}|$ is also plotted against V_{in} in Figure 20(d) showing that both transistors are momentarily ON as Vin passes through voltages between GND and VDD, resulting in a pulse of current drawn from the power supply.

The operation of the cmos inverter can be divided into five regions. The state of each transistor in each region is shown in Table 3.In region A the nMOS transistor is OFF so the pMOS transistor pulls the output to V_{DD}. In region B, the nMOS transistor starts to turn ON, pulling the output down. In region C, both transistors are in saturation. Notice that ideal transistors are only in region C for $V_{in}=V_{DD}/2$.

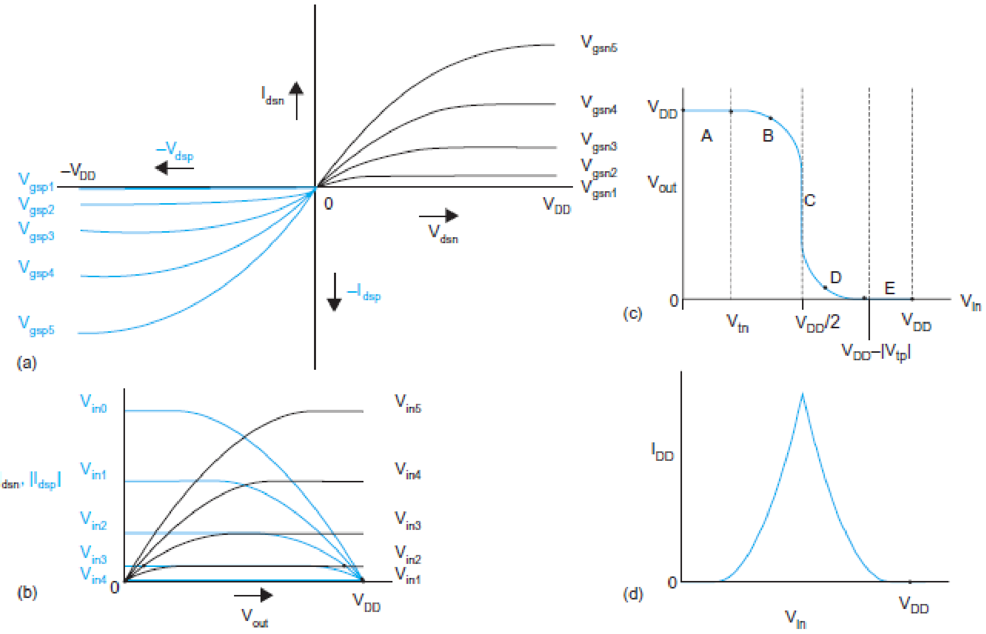
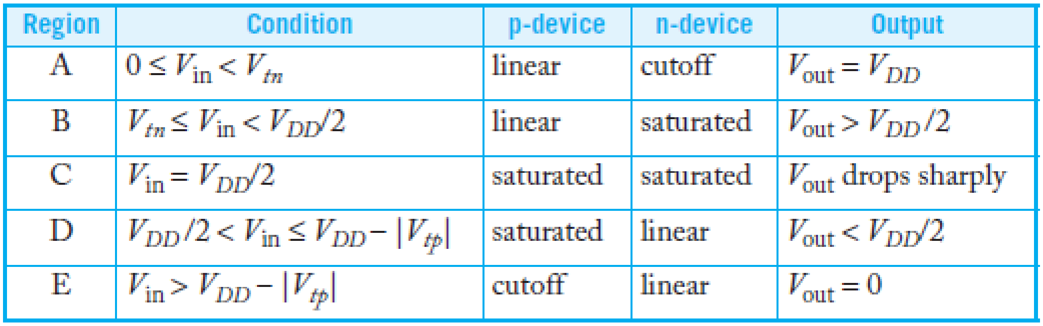


Fig.20. Graphical derivation of CMOS inverter DC characteristic

In region D, the pMOS transistor is partially ON and in region E, it is completely OFF, leaving the nMOS transistor to pull the output down to ground. Also the inverter's current consumption is zero when the input is within a threshold voltage of the VDD or GND rails. This feature is important for low power operation.

Table.3. Summary of CMOS inverter operation



Region	Condition	p-device	n-device	Output
A	$0 \le V_{in} < V_{tn}$	linear	cutoff	$V_{out} = V_{DD}$
B	$V_{tn} \le V_{in} < V_{DD}/2$	linear	saturated	$V_{out} > V_{DD}/2$
C	$V_{in} = V_{DD}/2$	saturated	saturated	V_{out} drops sharply
D	$V_{DD}/2 < V_{in} \le V_{DD} - \lvert V_{tp} \rvert$	saturated	linear	$V_{out} < V_{DD}/2$
E	$V_{in} > V_{DD} - \lvert V_{tp} \rvert$	cutoff	linear	$V_{out} = 0$

The given fig.21. Shows that the pMOS transistor is twice as wide as the nMOS transistor to achieve approximately equal betas. Simulation matches the simple models reasonably well, although the transition is not quite as steep because transistors are not ideal current sources in saturation. The crossover point where $V_{inv} = V_{in} = V_{out}$ is called the input threshold. Because

both mobility and the magnitude of the threshold voltage decrease with temperature for nMOS and pMOS transistors, the input threshold of the gate is only weakly sensitive to temperature.

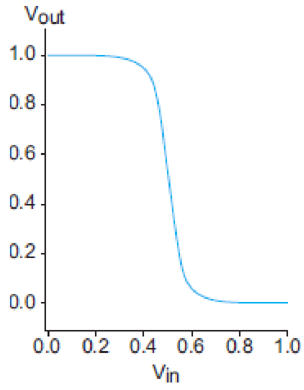


Fig.21. Simulated CMOS inverter DC characteristic

Beta Ratio Effects:

We have seen that for βp /βn, the inverter threshold voltage V_{inv} is $V_{DD}/2$. This may be desirable because it maximizes noise margins and allows a capacitive load to charge and discharge in equal times by providing equal current source and sink capabilities. Inverters with different beta ratios r =βp / βn are called skewed inverters. If r >1, the inverter is HI-skewed. If r <1, the inverter is LO-skewed. If r =1, the inverter has normal skew or is unskewed. A HI-skew inverter has a stronger pMOS transistor.

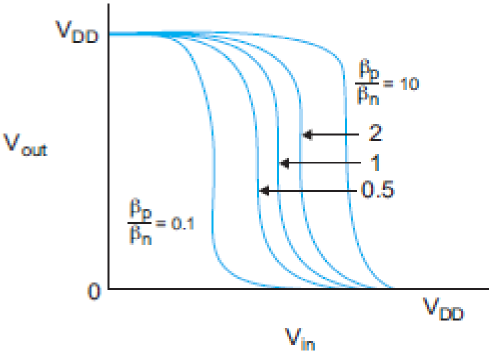


Fig.22. Transfer characteristics of skewed inverters

Therefore, if the input is VDD /2, we would expect the output will be greater than VDD /2. In other words, the input threshold must be higher than for an unskewed inverter. Similarly, a LO-skew inverter has a weaker pMOS transistor and thus a lower switching threshold.Figure.22 explores the impact of skewing the beta ratio on the DC transfer characteristics. As the beta ratio is changed, the switching threshold moves. However, the output voltage transition remains sharp. Gates are usually skewed by adjusting the widths of transistors while maintaining minimum length for speed.

DC transfer characteristics of other static CMOS gates can be understood by collapsing the gates into an equivalent inverter. Series transistors can be viewed as a single transistor of greater length. If only one of several parallel transistors is ON, the other transistors can be ignored. If several parallel transistors are ON, the collection can be viewed as a single transistor of greater width.

Noise Margin:

Noise margin is closely related to the DC voltage characteristics. This parameter allows you to determine the allowable noise voltage on the input of a gate so that the output will not be corrupted. The specification most commonly used to describe noise margin (or noise immunity) uses two parameters: the LOW noise margin, NML, and the HIGH noise margin, NMH. With reference to Figure 23. NML is defined as the difference in maximum LOW input voltage recognized by the receiving gate and the maximum LOW output voltage produced by the driving gate.

$$NM_L = V_{IL} - V_{OL}$$

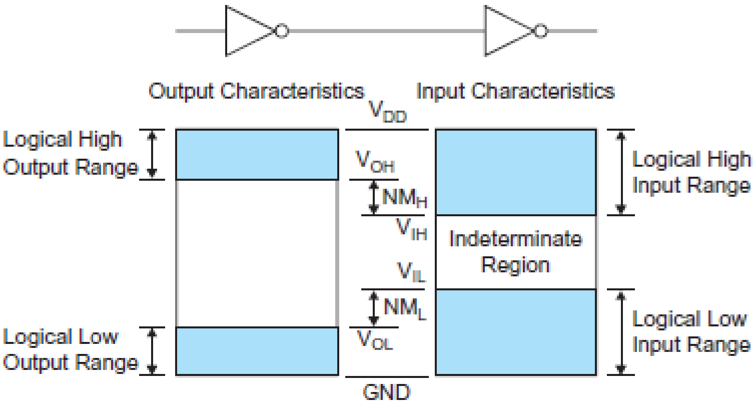


Fig.23. Noise margin definitions

The value of NMH is the difference between the minimum HIGH output voltage of the driving gate and the minimum HIGH input voltage recognized by the receiving gate. Thus,

$$NM_H = V_{OH} - V_{IH}$$

VOH=minimum HIGH output voltage
VOL =maximum LOW output voltage
VIH =minimum HIGH input voltage
VIL =maximum LOW input voltage

Inputs between VIL and VIH are said to be in the indeterminate region or forbidden zone and do not represent legal digital logic levels. Therefore, it is generally desirable to have VIH as close as possible to VIL and for this value to be midway in the "logic swing," VOL to VOH. This implies that the transfer characteristic should switch abruptly; that is, there should be high gain in the transition region.

Note that the output is slightly degraded when the input is at its worst legal value; this is called noise feed through or propagated noise. DC analysis gives us the static noise margins specifying the level of noise that a gate may see for an indefinite duration. Larger noise pulses may be acceptable if they are brief; these are described by dynamic noise margins specified by a maximum amplitude as a function of the duration. Unfortunately, there is no simple amplitude-duration product that conveniently specifies dynamic noise margins.

Ratioed Inverter Transfer Characteristics:

A more practical circuit called a pseudo-Nmos inverter is shown in fig.6.,It is uses a pMOS transistor pull-up or load that has its gate permanently grounded to approximate a constant current source. Pseudo-nMOS circuit get their name from the early nMOS technology in which only nMOS transistors were available; the grounded pMOS transistor is reminiscent of a depletion mode n MOS transistor that is always ON. The transfer characteristics may again be derived by finding V_{out} for a given V_{in}.

Pass Transistor DC Characteristics:

The nMOS transistors pass ‘0’s well but 1s poorly. We are now ready to better define “poorly.” Figure. 24.,(a) shows an nMOS transistor with the gate and drain tied to V_{DD}. Imagine that the source is initially at $V_s = 0$. $V_{gs} > V_{tn}$, so the transistor is ON and current flows. If the voltage on the source rises to $V_s = V_{DD} - V_{tn}$, V_{gs} falls to V_{tn} and the transistor cuts itself OFF. Therefore, nMOS transistors attempting to pass a 1 never pull the source above $V_{DD} - V_{tn}$. This loss is sometimes called a threshold drop.

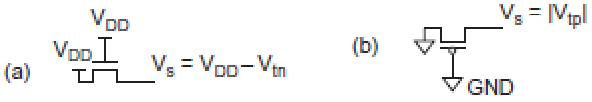


Fig.24.Pass transistor threshold drop

Moreover, when the source of the nMOS transistor rises, V_{sb} becomes nonzero. pMOS transistors pass 1s well but 0s poorly. If the pMOS source drops below $|V_{tp}|$, the transistor cuts off. Hence, pMOS transistors only pull down to within a threshold above GND, as shown in Figure 24.(b).

Tristste Inverter:

By cascading a transmission gate with an inverter, the tristate inverter constructed. When EN =0 and ENb=1,the output of the inverter is in a tristate condition (the Y output is not driven by the A output).When EN=1 and ENb=0,the Y output is equal to the complement of A.

1.3.5..Device models :

SPICE provides a wide variety of MOS transistor models with various trade-offs between complexity and accuracy. Level 1 and Level 3 models were historically important, but they are no longer adequate to accurately model very small modern transistors. BSIM models are more accurate and are presently the most widely used.

Level 1 models:

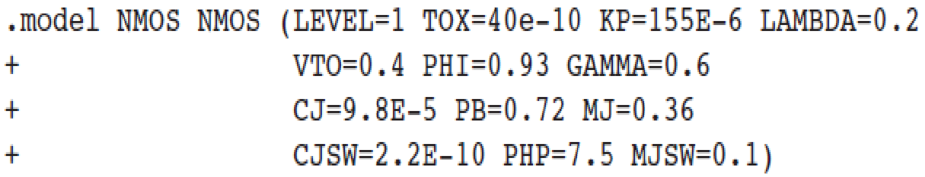
$$I_{ds} = \begin{cases} 0 & V_{gs} < V_t & \text{cutoff} \\ \text{KP}\dfrac{W_{\text{eff}}}{L_{\text{eff}}}\left(1+\text{LAMBDA}\times V_{ds}\right)\left(V_{gs}-V_t-{V_{ds}}/{2}\right)V_{ds} & V_{ds} < V_{gs}-V_t & \text{linear} \\ \dfrac{\text{KP}}{2}\dfrac{W_{\text{eff}}}{L_{\text{eff}}}\left(1+\text{LAMBDA}\times V_{ds}\right)\left(V_{gs}-V_t\right)^2 & V_{ds} > V_{gs}-V_t & \text{saturation} \end{cases}$$

The basic current models are given as

The threshold voltage is modulated by the source-to-body voltage Vsb through the body effect. For nonnegative Vsb , the threshold voltage is

$$V_t = \text{VTO} + \text{GAMMA}\left(\sqrt{\text{PHI}+V_{sb}} - \sqrt{\text{PHI}}\right)$$

Level 1 models are useful for teaching because they are easy to correlate with hand analysis, but are too simplistic for modern design. Figure 9.1 gives an example of a Level 1 model illustrating the syntax. The model also includes terms to compute the diffusion Capacitance.



```
.model NMOS NMOS (LEVEL=1 TOX=40e-10 KP=155E-6 LAMBDA=0.2
+                 VTO=0.4 PHI=0.93 GAMMA=0.6
+                 CJ=9.8E-5 PB=0.72 MJ=0.36
+                 CJSW=2.2E-10 PHP=7.5 MJSW=0.1)
```

Fig 9.1 Sample Level 1 Model

Level 2 and 3 Models:

The SPICE Level 2 and 3 models add effects of velocity saturation, mobility degradation, sub threshold conduction, and drain-induced barrier lowering. The Level 2 model is based on the Grove-Frohman equations [Frohman69], while the Level 3 model is based on empirical equations that provide similar accuracy, faster simulation times, and better convergence. However, these models still do not provide good fits to the measured I-V characteristics of modern transistors.

BSIM Models:

The Berkeley Short-Channel IGFE Model (BSIM) is a very elaborate model that is now widely used in circuit simulation. The models are derived from the underlying device physics but use an enormous number of parameters to fit the behavior of modern transistors. BSIM versions 1, 2, 3v3, and 4 are implemented as SPICE levels 13, 39, 49, and 54,respectively.BSIM is quite good for digital circuit simulation. Features of the model include:

1. Continuous and differentiable I-V characteristics across subthreshold, linear and saturation regions.

2. Sensitivity of parameters.

3. Detailed threshold voltage model. It includes body effect.

4. Velocity saturation, mobility degradation and short channel effects.

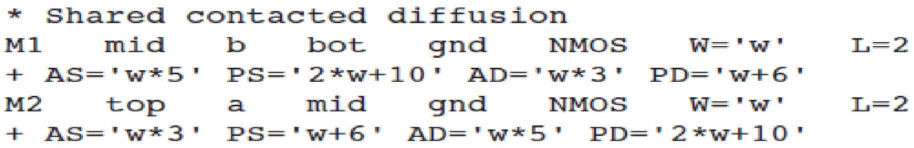
5. Diffusion capacitance and resistance models.

6. Multiple gate capacitance model.

In BSIM, different ranges of lengths and widths are specified by LMIN, LMAX, WMIN, WMAX.BSIM model is complicated one.

Diffusion capacitance model:

The p–n junction between the source or drain diffusion and the body forms a diode. We have seen that the diffusion capacitance determines the parasitic delay of a gate and depends on the area and perimeter of the diffusion.



```
* Shared contacted diffusion
M1    mid    b    bot    gnd    NMOS    W='w'    L=2
+ AS='w*5' PS='2*w+10' AD='w*3' PD='w+6'
M2    top    a    mid    gnd    NMOS    W='w'    L=2
+ AS='w*3' PS='w+6' AD='w*5' PD='2*w+10'
```

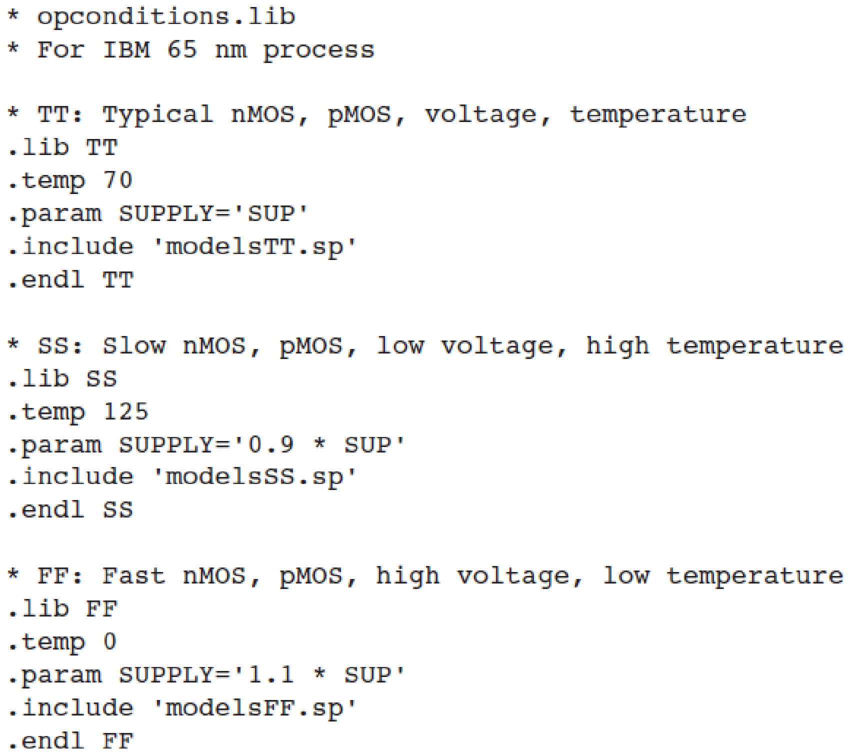
Fig 9.2 SPICE model of transistors with shared contacted diffusion

A SPICE description of the shared contacted diffusion case is shown in Figure 9.2.In the SPICE model, the diffusion capacitance between source and body is given by,

$$C_{sb} = \mathrm{AS} \times \mathrm{CJ} \times \left(1 + \frac{V_{sb}}{\mathrm{PB}}\right)^{-MJ} + \mathrm{PS} \times \mathrm{CJSW} \times \left(1 + \frac{V_{sb}}{\mathrm{PHP}}\right)^{-MJSW}$$

Design corners:

Engineers often simulate circuits in multiple design corners to verify operation across variations in device characteristics and environment. HSPICE includes the .lib statement that makes changing libraries easy. In the stimulus, the .alter statement is used to repeat the simulation with changes. In this case, the design corner is changed. The library file is given in Figure 9.3. Depending on what library was specified, the temperature is set (in degrees Celsius, with .temp) and the VDD value SUPPLY is calculated from the nominal SUP.



```
* opconditions.lib
* For IBM 65 nm process

* TT: Typical nMOS, pMOS, voltage, temperature
.lib TT
.temp 70
.param SUPPLY='SUP'
.include 'modelsTT.sp'
.endl TT

* SS: Slow nMOS, pMOS, low voltage, high temperature
.lib SS
.temp 125
.param SUPPLY='0.9 * SUP'
.include 'modelsSS.sp'
.endl SS

* FF: Fast nMOS, pMOS, high voltage, low temperature
.lib FF
.temp 0
.param SUPPLY='1.1 * SUP'
.include 'modelsFF.sp'
.endl FF
```

Fig 9.3 CORNER SPICE deck

1.4. Scaling:

The only constant in VLSI design is constant change. Feature size of the transistor has reduced by 30% every two to three years. As transistors become smaller, they switch faster, dissipate less power, and are cheaper to Manufacture. Designers need to be able to predict the effect of this feature size scaling on chip performance to plan future products, ensure existing products will scale gracefully to future processes for cost reduction, and anticipate looming design challenges.

1.4.1.Transistor Scaling:

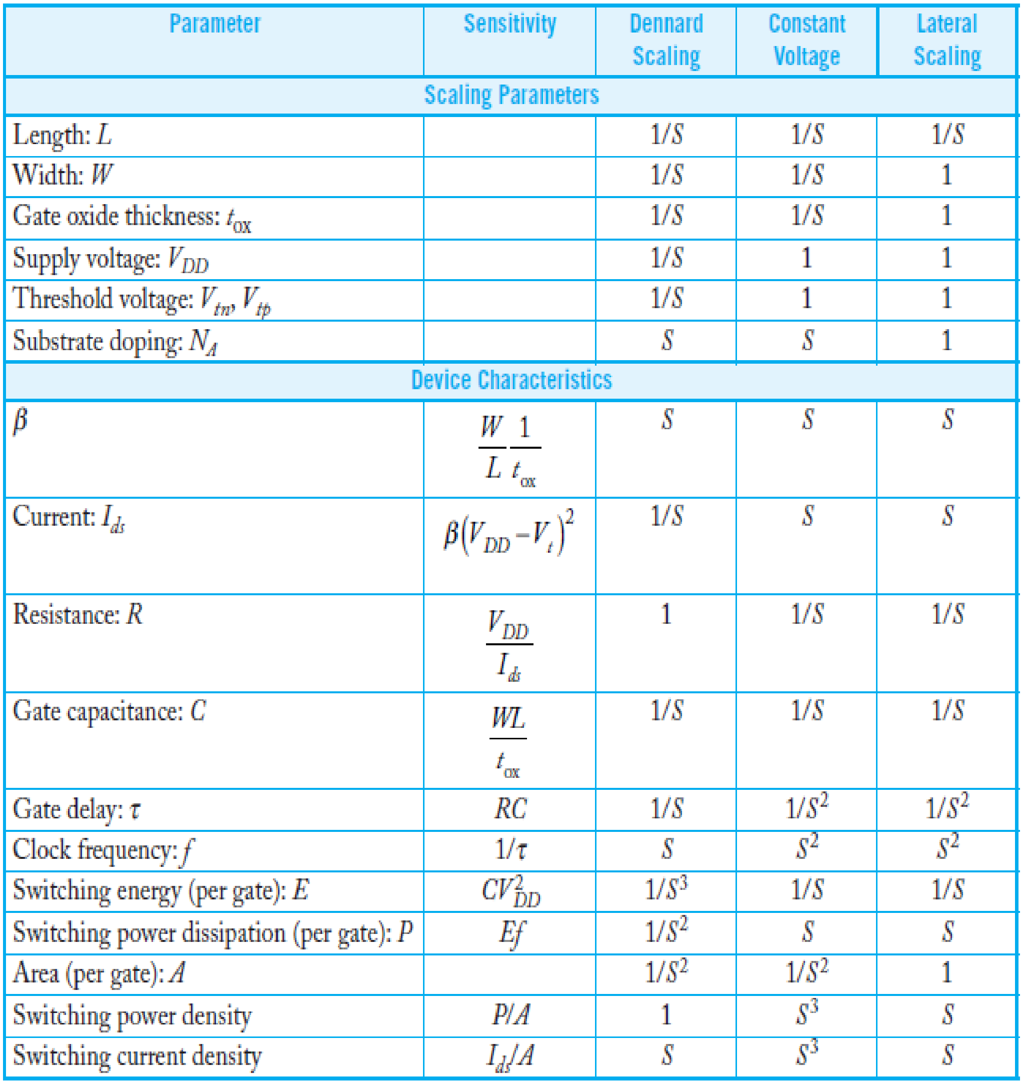
Dennard's Scaling Law predicts that the basic operational characteristics of a MOS transistor can be preserved and the performance improved if the critical parameters of a device are scaled by a dimensionless factor S. These parameters include the following:

- All dimensions (in the x, y, and z directions)
- Device voltages
- Doping concentration densities

This approach is also called constant field scaling because the electric fields remain the same as both voltage and distance shrink. In contrast, constant voltage scaling shrinks the devices but not the power supply. Another approach is lateral scaling, in which only the gate length is scaled. This is commonly called a gate shrink because it can be done easily to an existing mask database for a design.

The effects of these types of scaling are illustrated in Table 7.1. The industry generally scales process generations with $S = \sqrt{2}$; this is also called a 30% shrink. It reduces the cost (area) of a transistor by a factor of two. A 5% gate shrink (S =1.05) is commonly applied as a process becomes mature to boost the speed of components in that process.

TABLE 7.1 Influence of scaling on MOS device characteristics

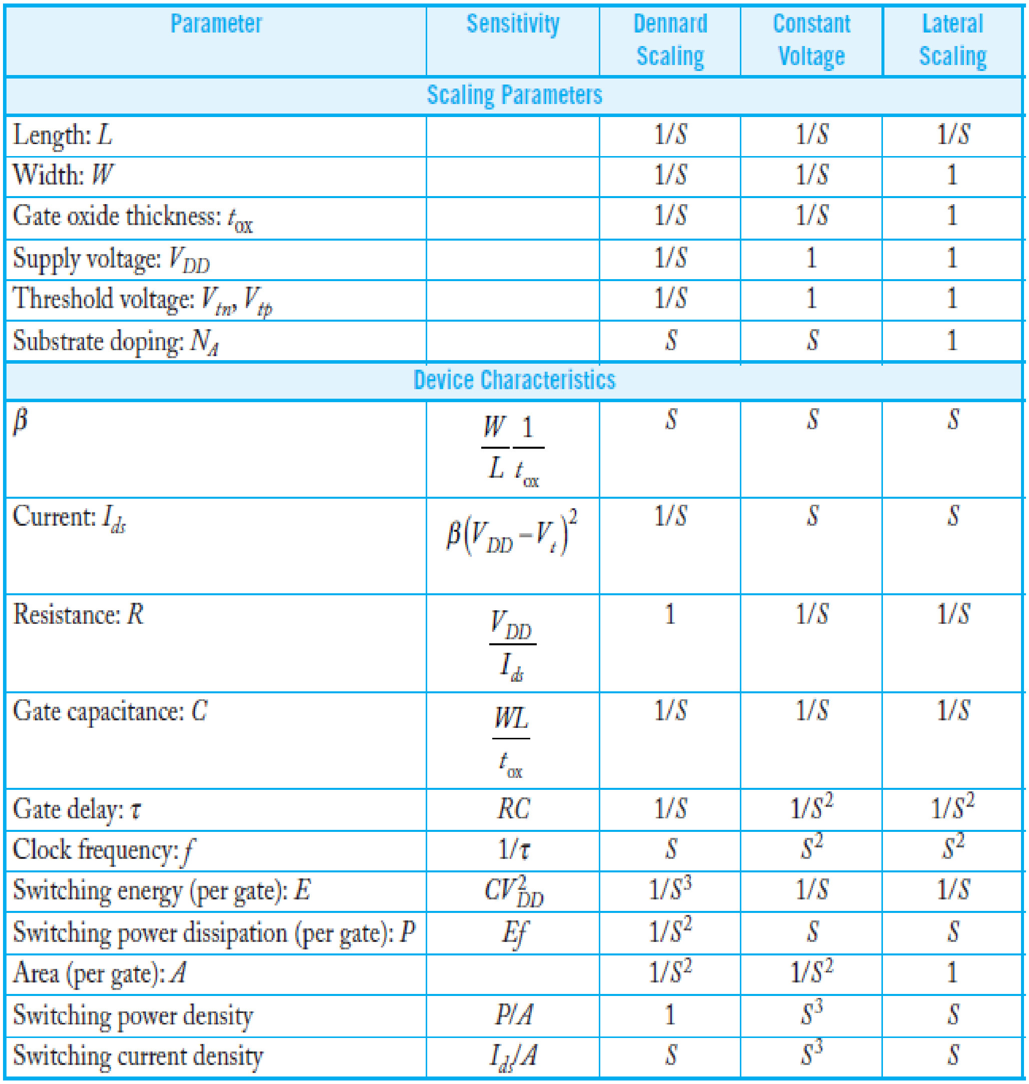


Parameter	Sensitivity	Dennard Scaling	Constant Voltage	Lateral Scaling
Scaling Parameters				
Length: L		$1/S$	$1/S$	$1/S$
Width: W		$1/S$	$1/S$	1
Gate oxide thickness: t_{ox}		$1/S$	$1/S$	1
Supply voltage: V_{DD}		$1/S$	1	1
Threshold voltage: V_{tn}, V_{tp}		$1/S$	1	1
Substrate doping: N_A		S	S	1
Device Characteristics				
β	$\frac{W}{L}\frac{1}{t_{ox}}$	S	S	S
Current: I_{ds}	$\beta\left(V_{DD}-V_t\right)^2$	$1/S$	S	S
Resistance: R	$\frac{V_{DD}}{I_{ds}}$	1	$1/S$	$1/S$
Gate capacitance: C	$\frac{WL}{t_{ox}}$	$1/S$	$1/S$	$1/S$
Gate delay: τ	RC	$1/S$	$1/S^2$	$1/S^2$
Clock frequency: f	$1/\tau$	S	S^2	S^2
Switching energy (per gate): E	CV_{DD}^2	$1/S^3$	$1/S$	$1/S$
Switching power dissipation (per gate): P	Ef	$1/S^2$	S	S
Area (per gate): A		$1/S^2$	$1/S^2$	1
Switching power density	P/A	1	S^3	S
Switching current density	I_{ds}/A	S	S^3	S

1.4.2.Interconnect Scaling:

Wires also tend to be scaled equally in width and thickness to maintain an aspect ratio close to 2. Table 7.2 shows the resistance, capacitance, and delay per unit length. Wires can be classified as local, semiglobal, and global. Local wires run within functional units and use the bottom layers of metal. Semiglobal (or scaled) wires run across larger blocks or cores, typically using middle layers of metal. Both local and semiglobal wires scale with feature size.

TABLE 7.2 Influence of scaling on interconnect characteristics



Parameter	Sensitivity	Dennard Scaling	Constant Voltage	Lateral Scaling
Scaling Parameters				
Length: L		$1/S$	$1/S$	$1/S$
Width: W		$1/S$	$1/S$	1
Gate oxide thickness: t_{ox}		$1/S$	$1/S$	1
Supply voltage: V_{DD}		$1/S$	1	1
Threshold voltage: V_{tn}, V_{tp}		$1/S$	1	1
Substrate doping: N_A		S	S	1
Device Characteristics				
β	$\frac{W}{L}\frac{1}{t_{ox}}$	S	S	S
Current: I_{ds}	$\beta\left(V_{DD}-V_t\right)^2$	$1/S$	S	S
Resistance: R	$\frac{V_{DD}}{I_{ds}}$	1	$1/S$	$1/S$
Gate capacitance: C	$\frac{WL}{t_{ox}}$	$1/S$	$1/S$	$1/S$
Gate delay: τ	RC	$1/S$	$1/S^2$	$1/S^2$
Clock frequency: f	$1/\tau$	S	S^2	S^2
Switching energy (per gate): E	CV_{DD}^2	$1/S^3$	$1/S$	$1/S$
Switching power dissipation (per gate): P	Ef	$1/S^2$	S	S
Area (per gate): A		$1/S^2$	$1/S^2$	1
Switching power density	P/A	1	S^3	S
Switching current density	I_{ds}/A	S	S^3	S

Global wires run across the entire chip using upper levels of metal. For example, global wires might connect cores to a shared cache. Global wires do not scale with feature size; indeed, they may get longer because die size has been gradually increasing.

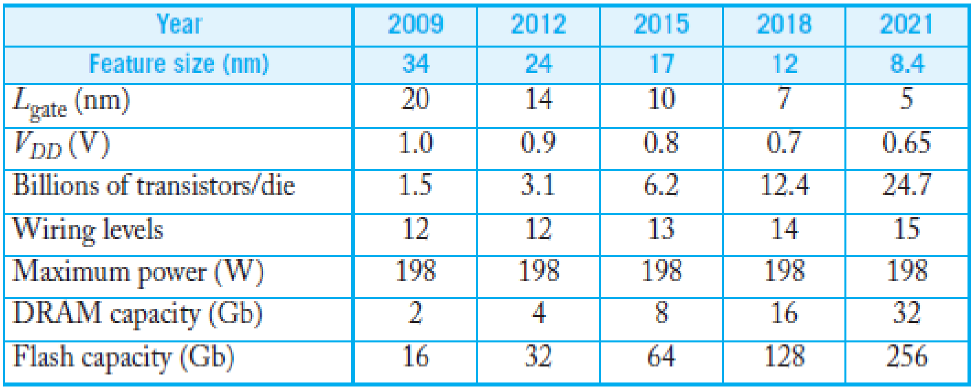
Most local wires are short enough that their resistance does not matter. Like gates, their capacitance per unit length is remaining constant, so their delay is improving just like gates. Semiglobal wires long enough to require repeaters are speeding up, but not as fast as gates. This is a relatively minor problem. Global wires, even with optimal repeaters, are getting slower as technology scales. The time to cross a chip in a nanometer process can be multiple cycles, and this delay must be accounted for in the microarchitecture.

1.4.3.International Technology Roadmap for Semiconductors:

The incredible pace of scaling requires cooperation among many companies and researchers both to develop compatible process steps and to anticipate and address future challenges before they hold up production. The Semiconductor Industry Association (SIA) develops and updates the International Technology Roadmap for Semiconductors (ITRS) to forge a consensus so that development efforts are not wasted on incompatible technologies and to predict future needs and direct research efforts. Such an effort to predict the future is inevitably prone to error, and the industry has scaled feature sizes and clock frequencies more rapidly than the roadmap predicted in the late 1990s.

The ITRS forecasts a major new technology generation, also called technology node, approximately every three years. Table 7.3 summarizes some of the predictions, particularly for high-performance microprocessors. However, serious challenges lie ahead, and major breakthroughs will be necessary in many areas to maintain the scaling on the roadmap.

TABLE 7.3 Predictions from the 2007 ITRS



Year	2009	2012	2015	2018	2021
Feature size (nm)	34	24	17	12	8.4
L_{gate} (nm)	20	14	10	7	5
V_{DD} (V)	1.0	0.9	0.8	0.7	0.65
Billions of transistors/die	1.5	3.1	6.2	12.4	24.7
Wiring levels	12	12	13	14	15
Maximum power (W)	198	198	198	198	198
DRAM capacity (Gb)	2	4	8	16	32
Flash capacity (Gb)	16	32	64	128	256

1.4.4.Impacts on Design:

One of the limitations of first-order scaling is that it gives the wrong impression of being able to scale proportionally to zero dimensions and zero voltage. In reality, a number of factors change significantly with scaling. This section attempts to peer into the crystal ball and predict some of the impacts on design for the future. These predictions are notoriously risky because chip designers have had an astonishing history of inventing ingenious solutions to seemingly insurmountable barriers.

1.4.5.Improved Performance and Cost:

The most positive impact of scaling is that performance and cost are steadily improving. Transistors are becoming cheaper each year, architects particularly need creative ideas of how to exploit growing numbers of transistors to deliver more or better functions.

1.4.6.Interconnect:

Scaled transistors are steadily improving in delay, but scaled global wires are holding the delay constant or getting worse. For short wires present inside logic gates, the wire RC delay is negligible. the "reachable radius" that a signal can travel in a cycle is getting smaller, as shown in Figure 7.1. For this, microarchitects are required to understand the floor plan and propose multiple pipeline stages for data to travel long distances across the die. Also, use more repeaters to reduce the delay.

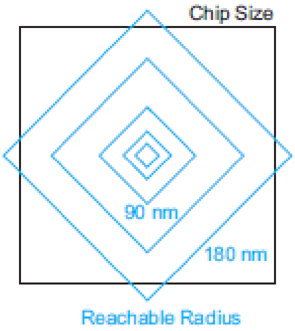


Fig 7.1 Reachable radius scaling

1.4.7.Power:

In classical constant field scaling, dynamic power density remains constant and overall chip power increases only slowly with die size. As clock frequency is increasing faster, the power density is increasing very rapidly. Subthreshold leakage power increased exponentially as threshold voltages decreased. Dynamic power consumption is increasing linearly because it will become uneconomical to cool the chip. Static power consumption is increasing rapidly for battery operated devices. Static power consumption caused by sub-threshold leakage was historically negligible. This is effective for threshold voltages below the range 0.3-0.4v.

1.4.8.Productivity:

The number of transistors that fit on a chip is increasing faster than designer productivity. One of the key tools to solve the productivity gap is design reuse.

1.4.9.Physical Limits:

Scaling cannot continue indefinitely because of the following reasons:

- Subthreshold leakage at low VDD and Vt
- Tunneling current through thin oxides
- Poor I-V characteristics due to DIBL and other short channel effects
- Dynamic power dissipation
- Lithography limitations
- Exponentially increasing costs of fabrication facilities and mask sets
- Electromigration
- Interconnect delay
- Variability

1.5.Propagation Delays

1.5.1..Delay Estimation:

In most designs there will be many logic paths that do not require any conscious effort when it comes to speed. These paths are already fast enough for the timing goals of the system. However, there will be a number of *critical paths* that limit the operating speed of the system and require attention to timing details. The critical paths can be affected at four main levels:

- The architectural/microarchitectural level
- The logic level
- The circuit level
- The layout level

The most leverage is achieved with a good microarchitecture. This requires a broad knowledge of both the algorithms that implement the function and the technology being targeted, such as how many gate delays fit in a clock cycle, how quickly addition occurs, how fast memories are accessed, and how long signals take to propagate along a wire. Trade-offs at the microarchitectural level include the number of pipeline stages, the number of execution units (parallelism), and the size of memories.

The next level of timing optimization comes at the logic level. Trade-offs include types of functional blocks (e.g., ripple carry vs. look ahead adders), the number of stages of gates in the clock cycle, and the fan-in and fan-out of the gates. The transformation from function to gates and registers can be done by experience, by experimentation, or, most often, by logic synthesis. Remember, however, that no amount of skillful logic design can overcome a poor micro architecture.

Once the logic has been selected, the delay can be tuned at the circuit level by choosing transistor sizes or using other styles of CMOS logic. Finally, delay is dependent on the layout. The floorplan (either manually or automatically generated) is of great importance because it determines the wire lengths that can dominate delay. Good cell layouts can also reduce parasitic capacitance.

Rise time t_r: It is the time for a waveform to rise from 10% to 90% of its steady state value

Fall time t_f: It is the time for a waveform to fall from 90% to 10% of its steady state value.

Delay time t_d: It is the time taken for a logic transition to pass from input to output.

Propagation delay time, t_{pd} : maximum time from the input crossing 50% to the output crossing 50%.

Contamination delay time, t_{cd} :minimum time from the input crossing 50% to the output crossing 50%.

RC delay model:

RC delay models approximate the nonlinear transistor I-V and C-V characteristics with an average resistance and capacitance over the switching range of the gate.

Effective Resistance:

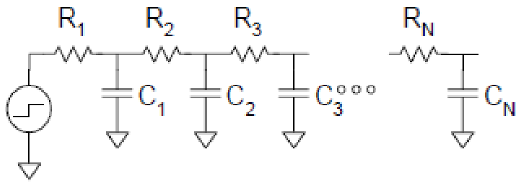
Unit width NMOS has resistance R, capacitance C. Unit width PMOS has resistance 2R,capacitance C. Capacitance proportional to width. Resistance is inversely proportional to width. When multiple transistors are in series, their resistance is sum of each individual resistance. When multiple transistors are in parallel, the resistance is lower if they all are ON.

Elmore delay model:

ON transistors are considered as resistors. Pull-up or pull-down networks are considered as RC ladders. Elmore Delay of a RC ladder:

$$t_{pd} = \Sigma R_{n-i} C_i$$

$$= R_1C_1 + (R_1 + R_2)C_2 + \ldots + (R_1 + R_2 + \ldots + R_N)C_N$$



Linear Delay Model:

The RC delay model showed that delay is a linear function of the fanout of a gate. In general, the normalized delay of a gate can be expressed in units of τ as

$$d = f + p$$

p is the *parasitic delay* inherent to the gate when no load is attached. '*f*' is the *effort delay* or *stage effort* that depends on the complexity and fanout of the gate:

$$f = gh$$

The complexity is represented by the *logical effort*, '*g*'. An inverter is defined to have a logical effort of 1. More complex gates have greater logical efforts, indicating that they take longer to drive a given fanout. For example, the logical effort of the 3-input NAND gate from the previous example is 5/3. A gate driving *h* identical copies of itself is said to have a *fanout* or *electrical effort* of *h*. If the load does not contain identical copies of the gate, the electrical effort can be computed as,

$$h = \frac{C_{out}}{C_{in}}$$

where C_{out} is the capacitance of the external load being driven and C_{in} is the input capacitance of the gate.

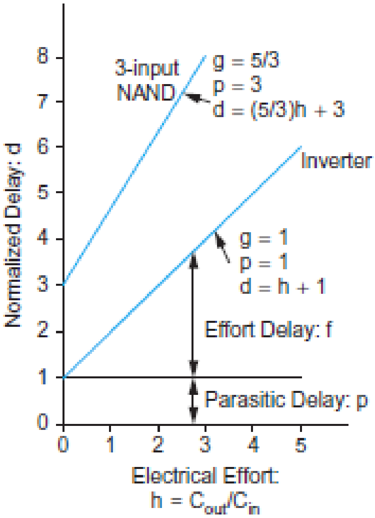
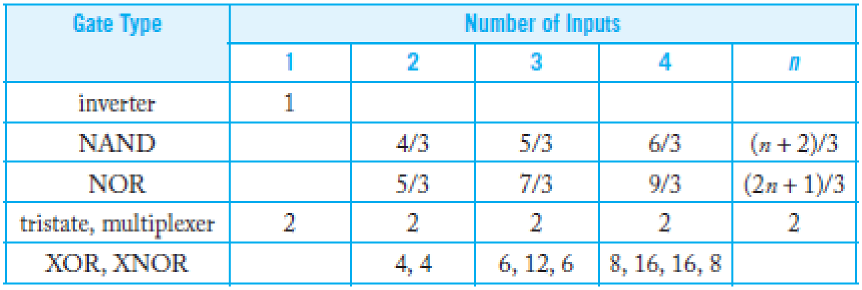


Fig.1.1. Normalized delay vs. fanout

Figure.1.1 plots normalized delay vs. electrical effort for an idealized inverter and 3-input NAND gate. The *y*-intercepts indicate the parasitic delay, i.e., the delay when the gate drives no load. The slope of the lines is the logical effort. The inverter has a slope of 1 by definition. The NAND has a slope of 5/3.

Logical Effort: Logical effort is the ratio of the input capacitance of a gate to the input capacitance of an inverter delivering the same output current. Logical effort can be measured in simulation from delay vs fanout plots as the ratio of the slope of the delay of the gate to the slope of the delay of an inverter. Table 1.1 lists the logical effort of common gates. The effort tends to increase with the number of inputs. NAND gates are better than NOR gates because the series transistors are nMOS rather than pMOS. Exclusive-OR gates are particularly costly and have different logical efforts for different inputs.

Table 1.1 Logical effort of common gates

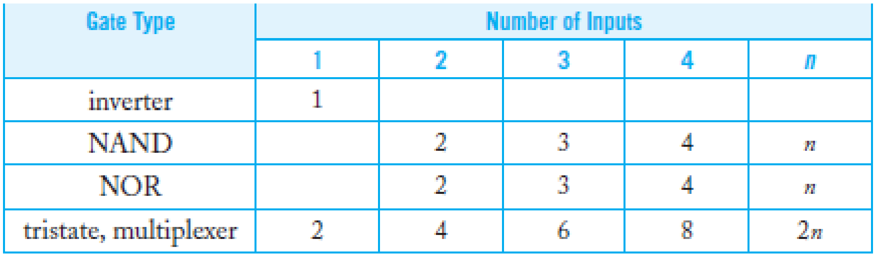


Gate Type	Number of Inputs				
	1	2	3	4	n
inverter	1				
NAND		4/3	5/3	6/3	$(n+2)/3$
NOR		5/3	7/3	9/3	$(2n+1)/3$
tristate, multiplexer	2	2	2	2	2
XOR, XNOR		4, 4	6, 12, 6	8, 16, 16, 8	

Parasitic Delay

The parasitic delay of a gate is the delay of the gate when it drives zero load. The parasitic delay also depends on the ratio of diffusion capacitance to gate capacitance. It can be estimated with RC delay models. A crude method good for hand calculations is to count only diffusion capacitance on the output node. Table 1.2 estimates the parasitic delay of common gates. Increasing transistor sizes reduces resistance but increases capacitance correspondingly, so parasitic delay is, on first order, independent of gate size.

Table 1.2 Parasitic delay of common gates



Gate Type	Number of Inputs				
	1	2	3	4	n
inverter	1				
NAND		2	3	4	n
NOR		2	3	4	n
tristate, multiplexer	2	4	6	8	$2n$

Effort delay:

It is the ratio of external load capacitance to input capacitance and thus changes with transistor widths. The capacitor ratio is called the electrical effort or fanout and the term indicating gate complexity is called the logical effort.

1.5.2.Logical effort and transistor sizing:

Designers often need to choose the fastest circuit topology and gate sizes for a particular logic function and to estimate the delay of the design. The method of Logical Effort provides a simple method “on the back of an envelope” to choose the best topology and number of stages of logic for a function. Based on the linear delay model, it allows the designer to quickly estimate the best number of stages for a path, the minimum possible delay for the given topology, and the gate sizes that achieve this delay. The techniques of Logical Effort will be revisited throughout this text to understand the delay of many types of circuits.

Delay in Multistage Logic Networks:

Figure 2.1 shows the logical and electrical efforts of each stage in a multistage path as a function of the sizes of each stage. The path of interest (the only path in this case) is marked with the dashed blue line. Observe that logical effort is independent of size, while electrical effort depends on sizes. This section develops some metrics for the path as a whole that are independent of sizing decisions.

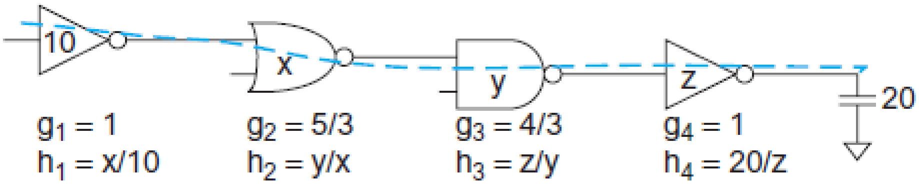


Fig 2.1 Multistage logic network

The path logical effort G can be expressed as the products of the logical effort of each stage along the path.

$$G=\prod g_i$$

The path electrical effort H can be given as the ratio of output capacitance,

$$H=C_{out(path)}/C_{in(path)}$$

The path effort F is the product of the stage efforts of each stage,

$$F=\prod f_i=\prod g_i h_i$$

In path that branch, F≠GH This is illustrated in Figure 2.2, a circuit with a two way branch. Consider a path from the primary input to one of the outputs. The path logical effort is G =1 ×1 =1 The path electrical effort is H =90/5 =18. Thus, GH =18. But $F =f_1 f_2=g_1h_1g_2h_2$ =1 ×6 ×1 ×6 =36. In other words, F =2GH in this path on account of the two-way branch.

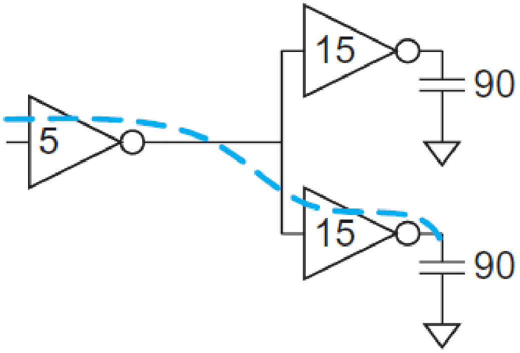


Fig 2.2 Circuit with two-way branch

We must introduce a new kind of effort to account for branching between stages of a path. This *branching effort b* is the ratio of the total capacitance seen by a stage to the capacitance on the path; in Figure 2.2 it is (15 +15)/15 =2.The branching effort 'b' is the ratio of the total capacitance seen by a stage to the capacitance on the path. In above fig. b=(15+15)/15=2.

$$b = \frac{C_{onpat\ h} + C_{offpat\ h}}{C_{onpat\ h}}$$

The path branching effort B is the product of the branching effort between stages,

$$B=\prod b_i$$

Now we can define the path effort F as the product of the logical electrical and branching efforts of the path.

$$F=GBH$$

The path delay D is the sum of the delays of each stage. It can also be written as the sum of the path effort delay D_F and path parasitic delay P.

$$D=\sum d_i=D_F+P$$

$$D_F=\sum P_i$$

The product of the stage efforts is F , independent of gate sizes. The path effort delay is the sum of the stage efforts. The path delay is minimized when each stage bears the same effort, that effort must be

$$\hat{f}=g_ih_i=F^{1/N}$$

Thus, the minimum possible delay of an *N*-stage path with path effort *F* and path parasitic delay *P* is,

$$D = NF^{1/N} + P$$

This is a key result of Logical Effort. It shows that the minimum delay of the path can be estimated knowing only the number of stages, path effort, and parasitic delays without the need to assign transistor sizes. This is superior to simulation, in which delay depends on sizes and you never achieve certainty that the sizes selected are those that offer minimum delay.

Choosing the best number of stages:

The logical effort cells us that NANDs are better than NORs and that gates with few inputs are better than gate with many.

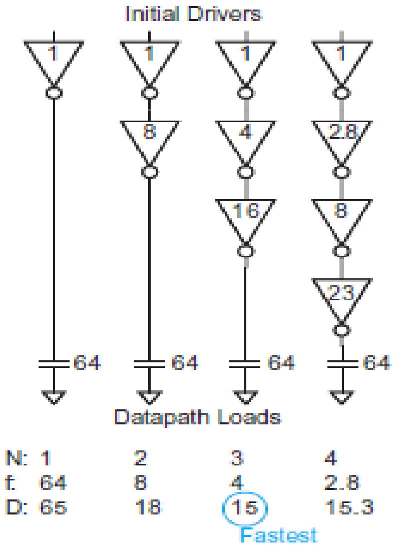


Fig 2.3 Comparison of different number of stages of buffers

Logical designers sometimes estimate delay by counting the number of stages of logic, assuming each stage has a constant “gate delay”. This is potentially misleading because it implies that the fastest circuits are those that use the fewest stages of logic of course the gate delay actually depends on the electrical effort. So sometimes using fewer stages results in more delay. For example, the 3 stage design is fastest and is much superior to single stage as shown in figure.2.3.

In general, you can always add inverters to the end of a path without changing its function The above fig. has n_1 stages and a path effort of F. Consider adding N-n_1 stages. The extra inverters do not change the path logical effort but do add parasitic delay. The delay of the new path is,

$$D=NF^{1/N}+\sum_{i=1}^{n1} P_i + (N-n_1)P_{inv}$$

Differentiating with respect to *N* and setting to 0 allows us to solve for the best number of stages, which we will call . The result can be expressed more compactly by defining

$$\rho = F^{\frac{1}{\hat{N}}}$$

Figure 2.4 plots the delay increase using a particular number of stages against the total number of stages, for p_{inv}=1. The *x*-axis plots the ratio of the actual number of stages to the ideal number. The *y*-axis plots the ratio of the actual delay to the best achievable. The curve is flat around the optimum.

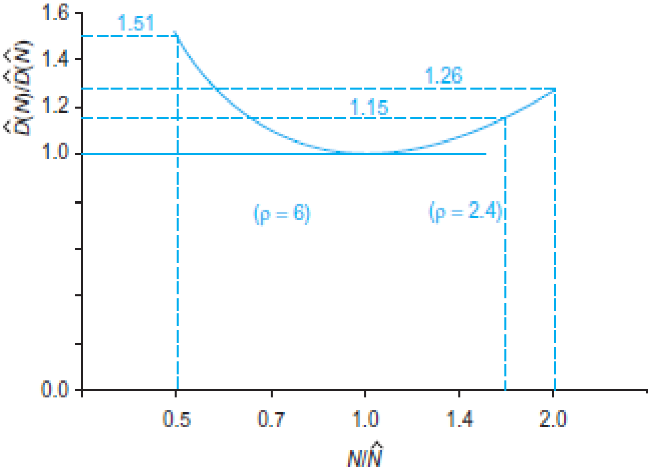


Fig 2.4 Sensitivity of delay to number of stages

1.6.Stick Diagrams

Because layout is time-consuming, designers need fast ways to plan cells and estimate area before committing to a full layout. Stick diagrams are easy to draw because they do not need to be drawn to scale. Figure 1.43 and the inside front cover show stick diagrams for an inverter and a 3-input NAND gate. While this book uses stipple patterns, layout designers use dry-erase markers or colored pencils. With practice, it is easy to estimate the area of a layout from the corresponding stick diagram even though the diagram is not to scale. Although schematics focus on transistors, layout area is usually determined by the metal wires.

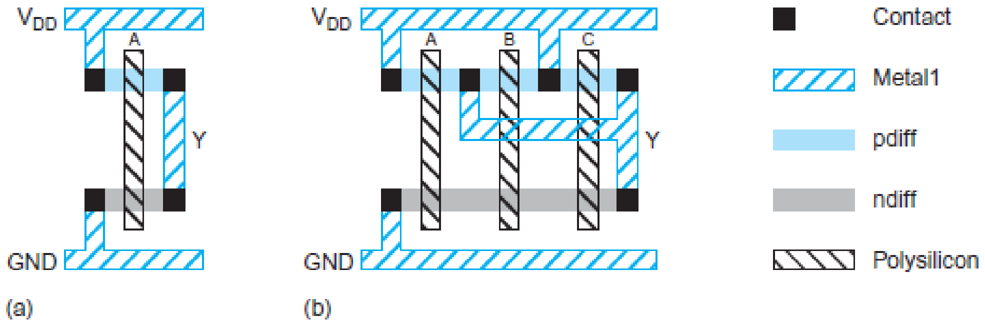


FIGURE 1.43 Stick diagrams of inverter and 3-input NAND gate. Color version on inside front cover.

Transistors are merely widgets that fit under the wires. We define a routing track as enough space to place a wire and the required spacing to the next wire. If our wires have a width of 4 λ and a spacing of 4λ to the next wire, the track pitch is 8 λ, as shown in Figure 1.44(a).

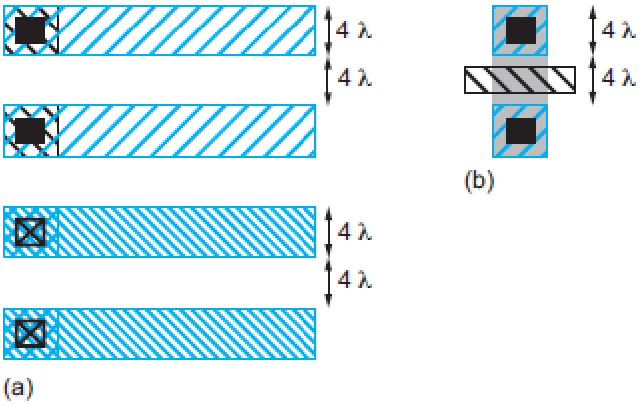


FIGURE 1.44 Pitch of routing tracks

This pitch also leaves room for a transistor to be placed between the wires (Figure 1.44(b)). Therefore, it is reasonable to estimate the height and width of a cell by counting the number of metal tracks and multiplying by 8 λ. A slight complication is the required spacing of 12 λ between nMOS and pMOS transistors set by the well, as shown in Figure 1.45(a).

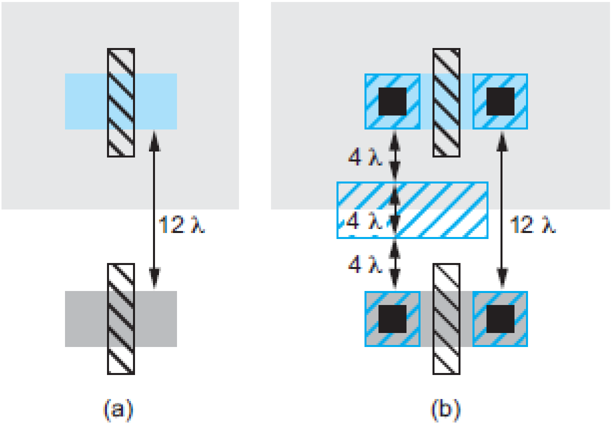


FIGURE 1.45 Spacing between nMOS and pMOS transistors

This space can be occupied by an additional track of wire, shown in Figure 1.45(b). Therefore, an extra track must be allocated between nMOS and pMOS transistors regardless of whether wire is actually used in that track.

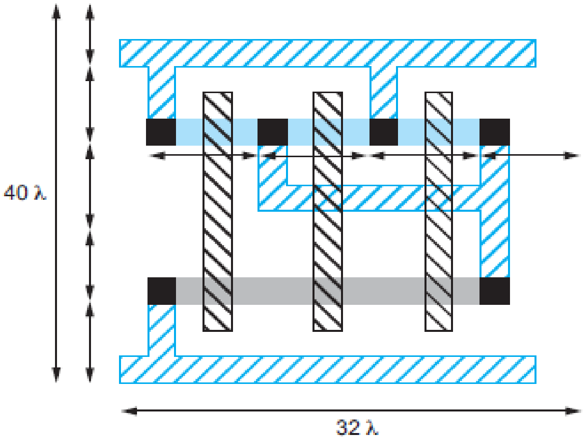


FIGURE 1.46 3-input NAND gate area estimation

Figure 1.46 shows how to count tracks to estimate the size of a 3-input NAND. There are four vertical wire tracks, multiplied by 8 λ per track to give a cell width of 32 λ. There are five horizontal tracks, giving a cell height of 40 λ. Even though the horizontal tracks are not drawn to scale, they are still easy to count. Figure 1.42 dimensions predicted by the stick diagram. If transistors are wider than 4 λ, the extra width must be factored into the area estimate. Of course, these estimates are oversimplifications of the complete design rules and a trial layout should be performed for truly critical cells.

Example 1.3

Sketch a stick diagram for a CMOS gate computing Y =$\overline{(A + B + C).D}$

SOLUTION: Figure shows a stick diagram.Counting horizontal and vertical pitches gives an estimated cell size of 40 by 48 λ

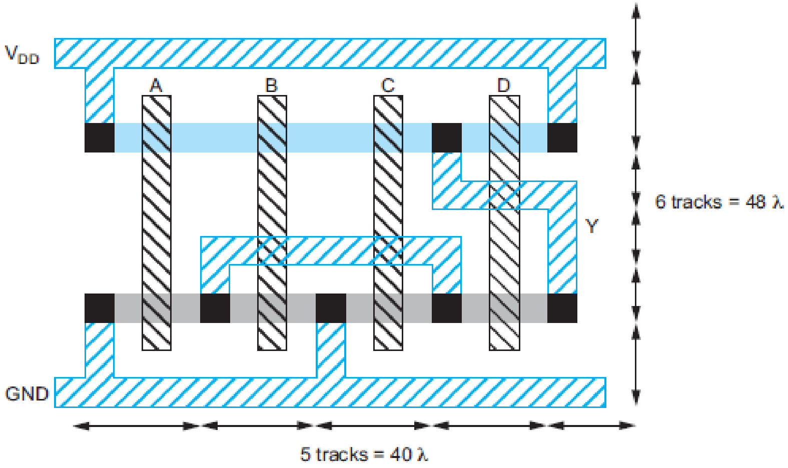


FIGURE 1.47 CMOS compound gate for function $Y = \overline{(A + B + C) \cdot D}$