VLSI DESIGN(Unit-2)

UNIT-II

Basic Circuit Concepts: Capacitance, resistance estimations- Sheet Resistance R_s, MOSDivice Capacitances, routing Capacitance, Analytic Inverter Delays, Driving large Capacitive Loads, Fan-in and fan-out.

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

Q)Define Sheet resistance of the MOS device.

SHEET RESISTANCE *Rs*

Consider a uniform slab of conducting material of resistivity ρ, of width *W*, thickness *t*, and
length between faces *L*. The arrangement is shown in Figure

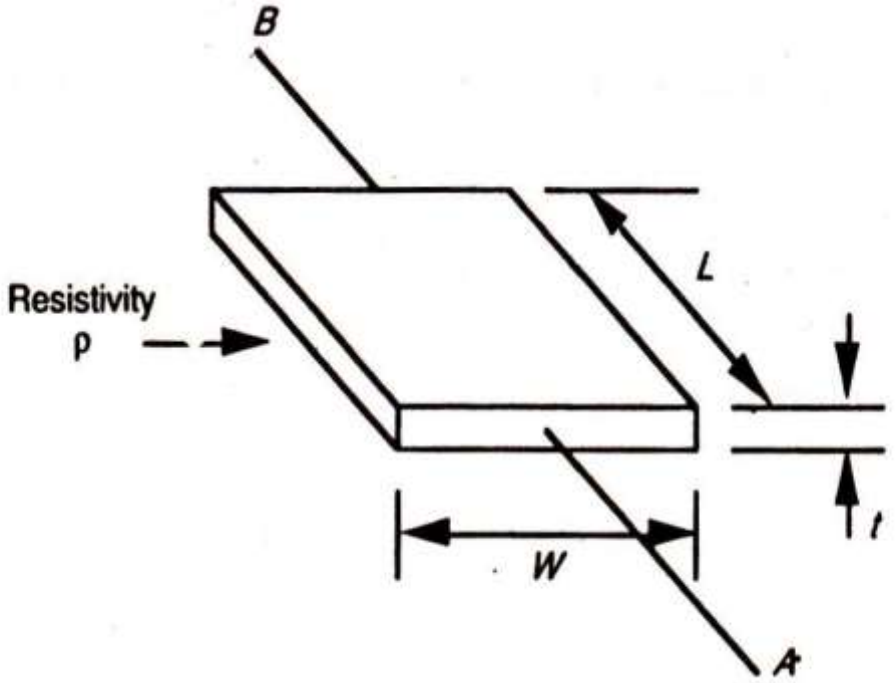


FIGURE Sheet resistance model .

With reference to Figure , consider the resistance R_{AB} between two opposite faces.

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

where

A = cross-section area

Thus

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

Now, consider the case in which $L = W$, that is, a square of resistive material, then

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

where

R_s = ohm per square or sheet resistance

Thus

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

For the MOS processes considered here, typical values of sheet resistance are given in Table

TABLE 4.1 Typical sheet resistances R_s of MOS layers for 5 μm*, and Orbit 2 μm* and 1.2 μm* technologies

Layer	*R_s ohm per square*		
	5 μm	*Orbit*	*Orbit 1.2 μm*
Metal	0.03	0.04	0.04
Diffusion (or active)**	10→50	20→45	20→45
Silicide	2→4	—	—
Polysilicon	15→100	15→30	15→30
n-transistor channel	$10^{4\dagger}$	$2 \times 10^{4\dagger}$	$2 \times 10^{4\dagger}$
p-transistor channel	$2.5 \times 10^{4\dagger}$	$4.5 \times 10^{4\dagger}$	$4.5 \times 10^{4\dagger}$

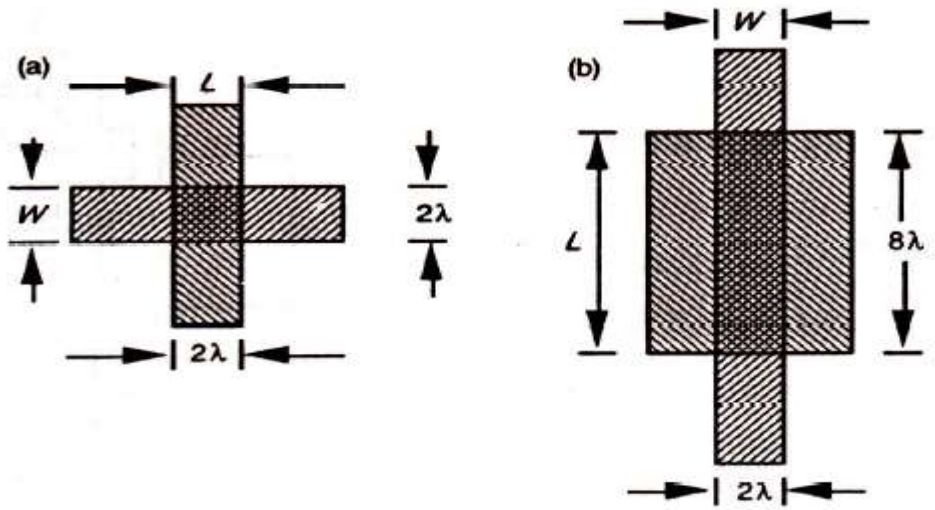


FIGURE Resistance calculation for transistor channels

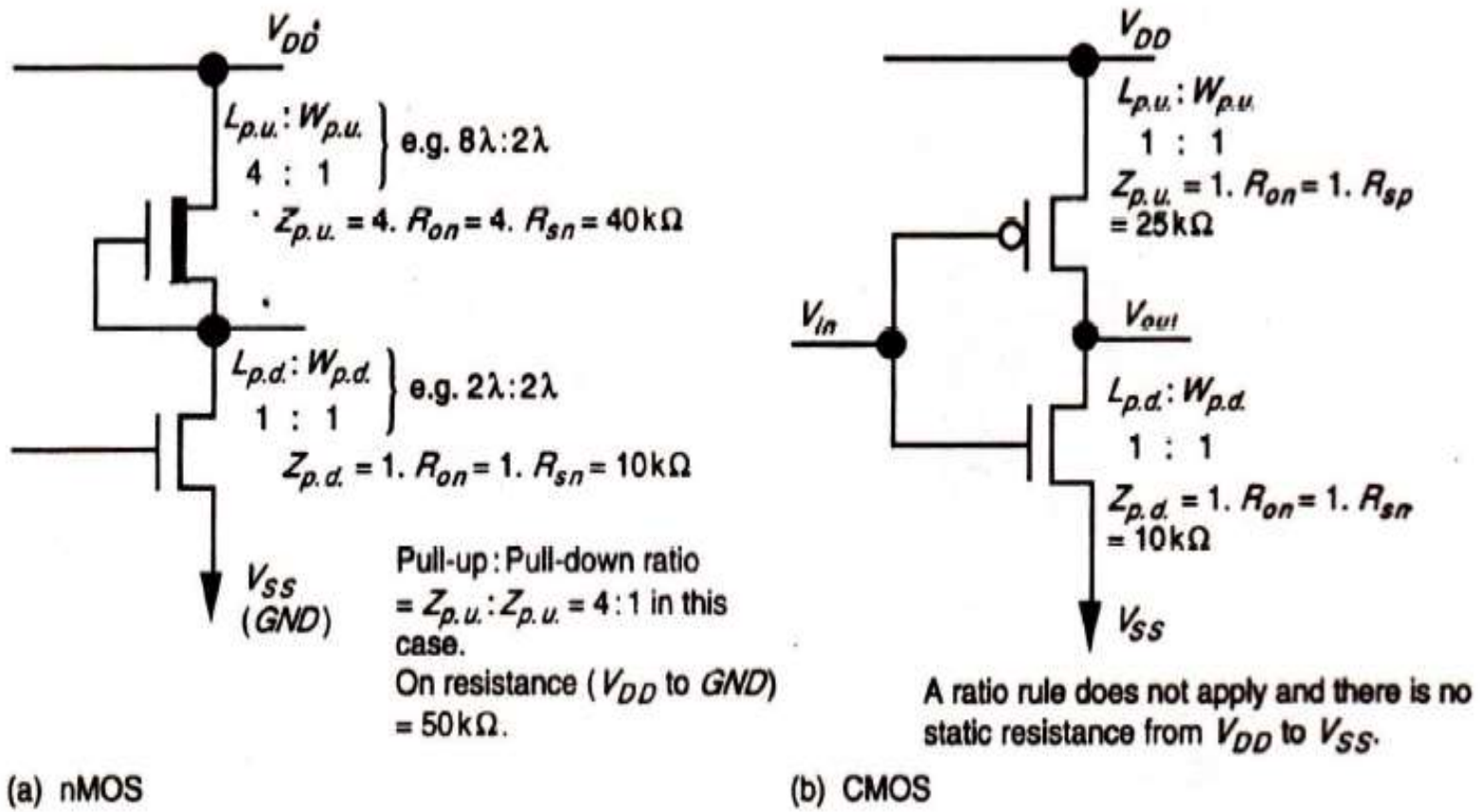
$$R = 1 \text{ square} \times R_s \frac{\text{ohm}}{\text{square}} = R_s = 10^4 \text{ ohm}^*$$

The length to width ratio, denoted Z, is 1:1 in this case. The transistor structure of Figure 4.2(b) has a channel length $L = 8\lambda$ and width $W = 2\lambda$. Therefore,

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

Thus, channel resistance

$$R = ZR_s = 4 \times 10^4 \text{ ohm}$$



Note: R_{on} = 'on' resistance; R_{sn} = n-channel sheet resistance; R_{sp} = p-channel sheet resistance.

FIGURE Inverter resistance calculation.

AREA CAPACITANCES OF LAYERS

For any layer, knowing the dielectric (silicon dioxide) thickness, we can calculate area capacitance as follows:

$$C = \frac{\varepsilon_0 \varepsilon_{ins} A}{D} \text{ farads}$$

where

D = thickness of silicon dioxide

A = area of plates

(and it is assumed that ε_0, A, and D are in compatible units, for example, ε_0 in farads/cm, A in cm^2, D in cm).

ε_{ins} = relative permittivity of $SiO_2 \div 4.0$

$\varepsilon_0 = 8.85 \times 10^{-14}$ F/cm (permittivity of free space)

A normal approach is to give layer area capacitances in $pF/\mu m^2$ (where μm = micron $= 10^{-6}$ meter $= 10^{-4}$ cm). The appropriate figure may be calculated as follows:

$$C\left(\frac{\text{pF}}{\mu\text{m}^2}\right) = \frac{\varepsilon_0 \varepsilon_{ins}}{D} \frac{\text{F}}{\text{cm}^2} \times \frac{10^{12}\text{pF}}{\text{F}} \times \frac{\text{cm}^2}{10^8\,\mu\text{m}^2}$$

(D in cm, ε_0 in farads/cm)

Typical values of area capacitance are set out in Table 4.2 for 5 μm technology and for Orbit 2 μm and 1.2 μm technologies.

TABLE Typical area capacitance values for MOS circuits
Capacitance

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

Notes: Relative value = specified value/gate to channel value for that technology.
*Poly. 1 and Poly. 2 are similar (also silicides where used).

STANDARD UNIT OF CAPACITANCE □ C_G

The unit is denoted □Cg and is defined the gate-to-channel capacitance of a MOS transistor having $W = L$ = feature size, that is, a 'standard' or 'feature size' square as in Figure

$\square C_g$ may be evaluated for any MOS process. For example, for 5 μm MOS circuits:

Area/standard square = 5 μm × 5 μm = 25 μm^2 (= area of minimum size transistor)
Capacitance value (from Table 4.2) = 4×10^{-4} pF/μm^2
Thus, standard value $\square C_g$ = 25 $\mu m^2 \times 4 \times 10^{-4}$ pF/μm^2 = .01 pF

or, for 2 μm MOS circuits (Orbit):

Area/standard square = 2 μm × 2 μm = 4 μm^2
Gate capacitance value (from Table 4.2) = 8×10^{-4} pF/μm^2
Thus, standard value $\square C_g$ = 4 $\mu m^2 \times 8 \times 10^{-4}$ pF/μm^2 = .0032 pF

and, for 1.2 μm MOS circuits (Orbit):

Area/standard square = 1.2 μm × 1.2 μm = 1.44 μm^2
Gate capacitance value (from Table 4.2) = 16×10^{-4} pF/μm^2
Thus, standard value $\square C_g$ =1.44 $\mu m^2 \times 16 \times 10^{-4}$ pF/μm^2 = .0023 pF

SOME AREA CAPACITANCE CALCULATIONS

Consider the area defined in Figure First, we must calculate the area relative to that of a standard gate.

$$\text{Relative area} = \frac{20\lambda \times 3\lambda}{2\lambda \times 2\lambda} = 15$$

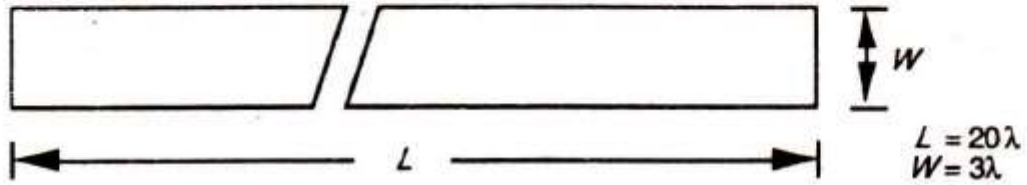


FIGURE Simple area for capacitance calculation

Now:

1. Consider the area in metal 1.
   Capacitance to substrate = relative area × relative C value
   $$= 15 \times 0.0750 \square C_g$$
   $$= 1.125 \square C_g$$
   That is, the defined area in metal has a capacitance to substrate 1.125 times that of a feature size square gate area.

2. Consider the same area in polysilicon.

   Capacitance to substrate = $15 \times 0.1 \square C_g$

   $= 1.5 \square C_g$

3. Consider the same area in n-type diffusion.

   Capacitance to substrate = $15 \times 0.25 \square C_g$

   $= 3.75 \square C_g^*$

Calculations of area capacitance values associated with structures occupying more than
one layer, as in Figure are equally straightforward

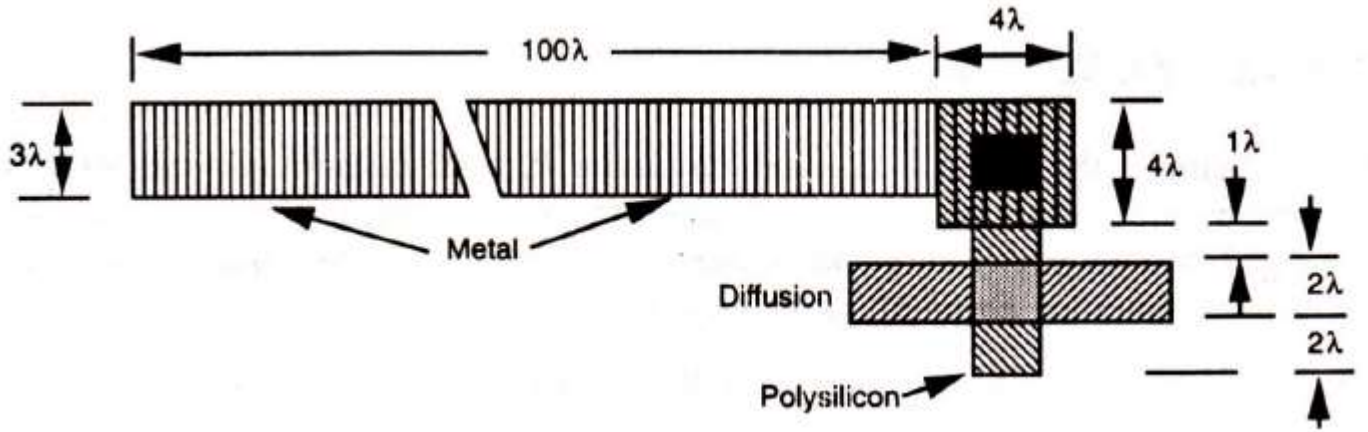


FIGURE Capacitance calculation (multilayer

Consider the metal area (less the contact region where the metal is connected to polysilicon and shielded from the substrate)

$$\text{Ratio} = \frac{\text{Metal area}}{\text{Standard gate area}} = \frac{100\lambda \times 3\lambda}{4\lambda^2} = 75$$

$$\text{Metal capacitance } C_m = 75 \times 0.075 = 5.625 \square C_g$$

Consider the polysilicon area (excluding the gate region)

$$\text{Polysilicon area} = 4\lambda \times 4\lambda + 3\lambda \times 2\lambda = 22\lambda^2$$

Therefore

$$\text{Polysilicon capacitance } C_p = \frac{22}{4} \times 0.1 = .55 \square C_g$$

For the transistor,

$$\text{Gate capacitance } C_g = 1 \square C_g$$

$$\text{Total capacitance } C_T = C_m + C_p + C_g \doteq 7.20 \square C_g$$

THE DELAY UNIT τ

We have developed the concept of sheet resistance *Rs* and standard gate capacitance unit

□ Cg If we consider the case of one standard (feature size square) gate area capacitance being charged through one feature size square of n channel resistance (that is, through *Rs* for an nMOS pass transistor channel), as in Figure we have:

Time constant $\tau = (1R_s$ (n channel) $\times 1\square C_g)$ seconds

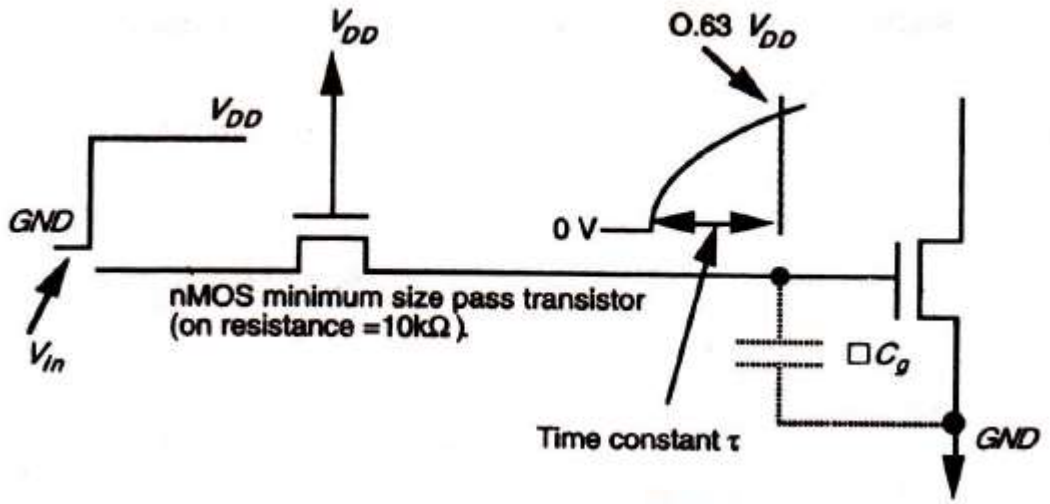


FIGURE Model for derivation of. τ

This can be evaluated for any technology and for 5 μm technology,

$$\tau = 10^4 \text{ ohm} \times 0.01 \text{ pF} = 0.1 \text{ nsec}$$

and for 2 μm (Orbit) technology,

$$\tau = 2 \times 10^4 \text{ ohm} \times 0.0032 \text{ pF} = 0.064 \text{ nsec}$$

and for 1.2 μm (Orbit) technology,

$$\tau = 2 \times 10^4 \text{ ohm} \times 0.0023 \text{ pF} = 0.046 \text{ nsec}$$

However, in practice, circuit wiring and parasitic capacitances must be allowed for so that the figure taken for τ is often increased by a factor of two or three so that for 5 μm circuit

τ = 0.2 to 0.3 nsec is a typical design figure used in assessing likely worst case delays.

Note that τ thus obtained is not much different from transit time τ_{sd} calculated from equation (2.2).

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

Note that V_{ds} varies as C_g charges from 0 volts to 63% of V_{DD} in period τ in Figure 4.6, so that an appropriate value for V_{ds} is the average value = 3 volts. For 5 μm technology, then,

$$\tau_{sd} = \frac{25\mu m^2 \text{ V sec}}{650 \text{ cm}^2 \text{ 3V}} \times \frac{10^9 \text{ nsec cm}^2}{10^8 \mu m^2}$$

$$= 0.13 \text{ nsec}$$

For 5 µm MOS technology $\tau = 0.3$ nsec is a very safe figure to use; and, for 2 µm Orbit MOS technology, $\tau = 0.2$ nsec is an equally safe figure to use; and, for 1.2 µm Orbit MOS technology, $\tau = 0.1$ nsec is also a safe figure.

INVERTER DELAYS

Consider the basic 4: 1 ratio nMOS inverter. In order to achieve the 4:1 *Zp.u.* to *Zp.d.* ratio, *Rp.u.* will be 4 *Rp.d.* and if *Rp.d.* is contributed by the minimum size transistor then, clearly, the resistance value associated with *Rp.u.* is

Rp.u. = $4R_S$ = 40 kQ Meanwhile, the *Rp.d.* value is *1Rs* = 10 kQ so that the delay associated with the inverter will depend on whether it is being turned on or off.

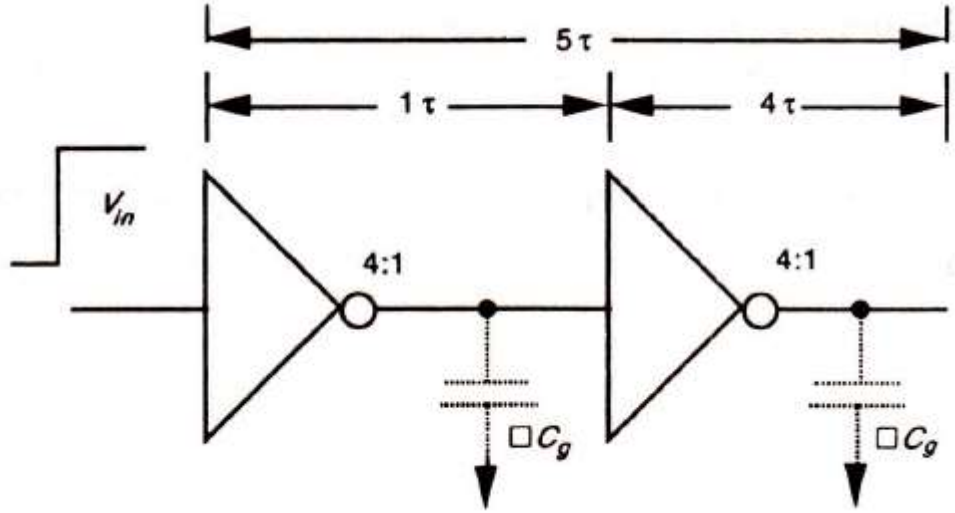


FIGURE nMOS Inverter pair delay.

through a pair of similar nMOS inverters is

$$T_d = (1 + Z_{p.u.}/Z_{p.d.})\tau$$

Thus, the inverter pair delay for inverters having 4:1 ratio is 5τ.

However, a single 4:1 inverter exhibits undesirable asymmetric delays since the delay in turning on is, for example, τ, while the corresponding delay in turning off is 4τ. Quite obviously, the asymmetry is worse when considering an inverter with an 8:1 ratio.

However, a single 4: 1 inverter exhibits undesirable asymmetric delays since the delay in turning on is, for example,τ, while the corresponding delay in turning off is 4τ. Quite obviously, the asymmetry is worse when considering an. inverter with an 8:1 ratio

.

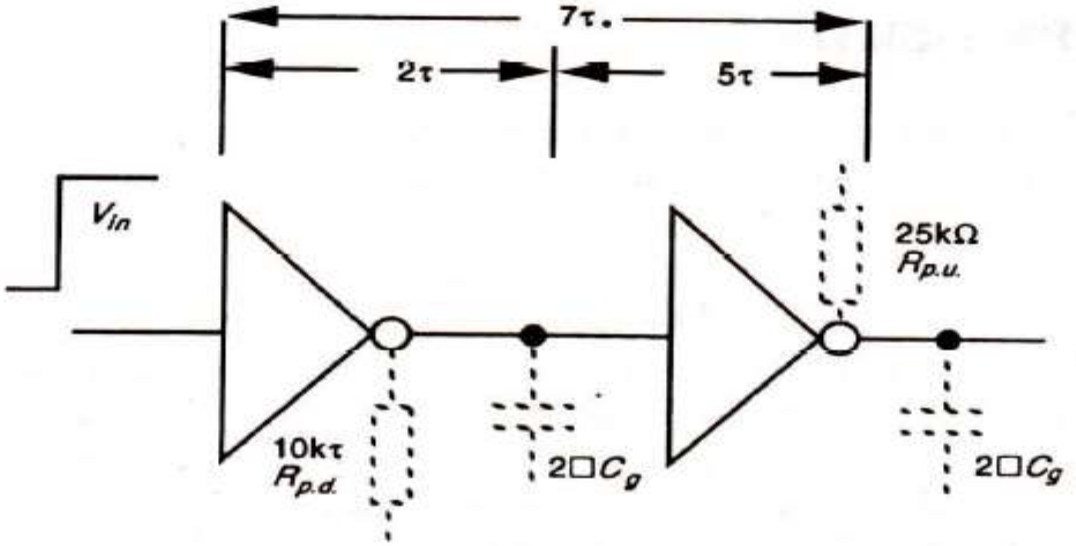


FIGURE Minimum size CMOS Inverter pair delay

A More Formal Estimation of CMOS Inverter Delay

A CMOS inverter, in general, either charges or discharges a capacitive load C_L and rise-time τ_r or fall-time τ_f can be estimated from the following simple analysis

The saturation current for the p-transistor is given by

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

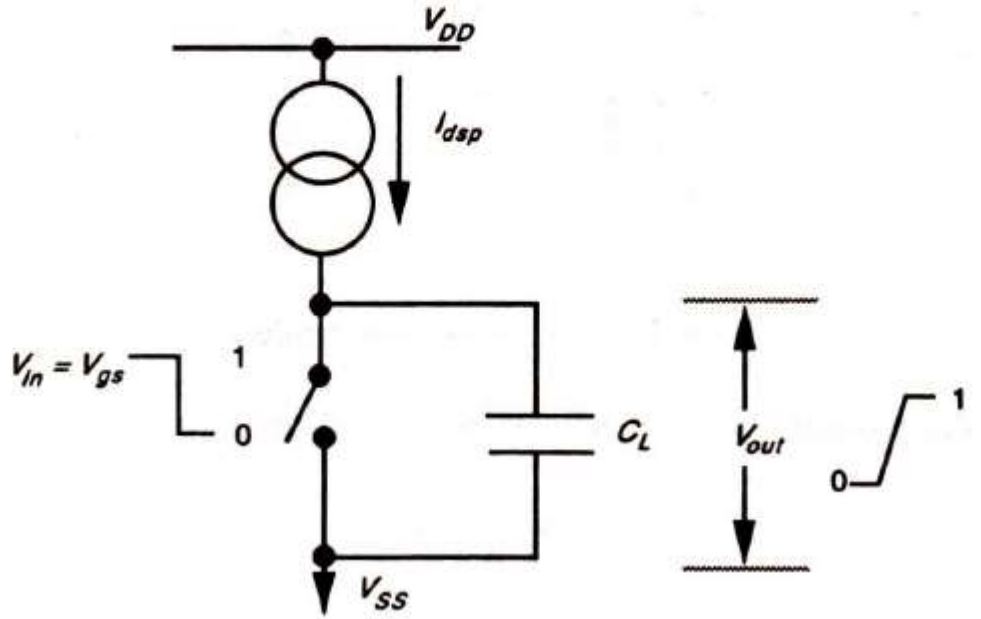


FIGURE Rise-time model

This current charges C_L and, since its magnitude is approximately constant, we have

$$V_{out} = \frac{I_{dsp}t}{C_L}$$

Substituting for I_{dsp} and rearranging we have

$$t = \frac{2C_L V_{out}}{\beta_p (V_{gs} - |V_{tp}|)^2}$$

We now assume that $t = \tau_r$ when $V_{out} = +V_{DD}$, so that

$$\tau_r = \frac{2V_{DD}C_L}{\beta_p (V_{DD} - |V_{tp}|)^2}$$

with $|V_{tp}| = 0.2V_{DD}$, then

$$\tau_r \doteq \frac{3C_L}{\beta_p V_{DD}}$$

This result compares reasonably well with a more detailed analysis in which the charging of C_L is divided, more correctly, into two parts: (1) saturation and (2) resistive region of the transistor.

Fall-time estimation

Similar reasoning can be applied to the discharge of C_L through the n-transistor. The circuit model in this case is given as Figure

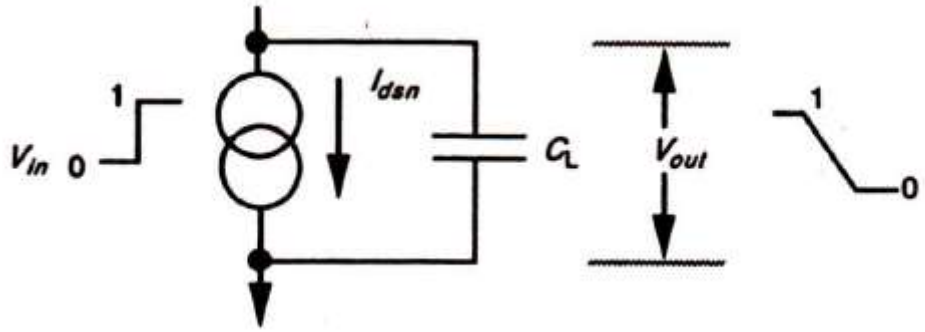


FIGURE Fall-time model

Making similar assumptions we may write for fall-time:

$$\tau_f \doteq \frac{3C_L}{\beta_n V_{DD}}$$

Summary of CMOS rise and fall factors

Using these expressions we may deduce that:

$$\frac{\tau_r}{\tau_f} = \frac{\beta_n}{\beta_p}$$

But $\mu_n = 2.5\ \mu_p$ and hence $\beta_n \div 2.5\beta_p$, so that the rise-time is slower by a factor of 2.5 when using minimum size devices for both 'n' and 'p'.

This simple model is quite adequate for most practical situations, but it should be recognized that it gives optimistic results. However, it does provide an insight into the factors which affect rise-times and fall-times as follows:

1. τ_r and τ_f are proportional to $1/V_{DD}$;
2. τ_r and τ_f are proportional to C_L;
3. $\tau_r = 2.5\tau_f$ for equal n- and p-transistor geometries.

DKMNG LARGE CAPACITIVE LOADS

The problem of driving comparatively large capacitive loads arises when signals must be propagated from the chip to off chip destinations . . Generally, typical off chip capacitances may be several orders higher than on chip □Cg values. For example, if the off chip load is denoted C_L then

$$C_L \geq 10^4 \square C_g \text{ (typically)}$$

Clearly capacitances of this order must be driven through low resistances, otherwise excessively long delays will occur.

Cascaded Inverters as Drivers

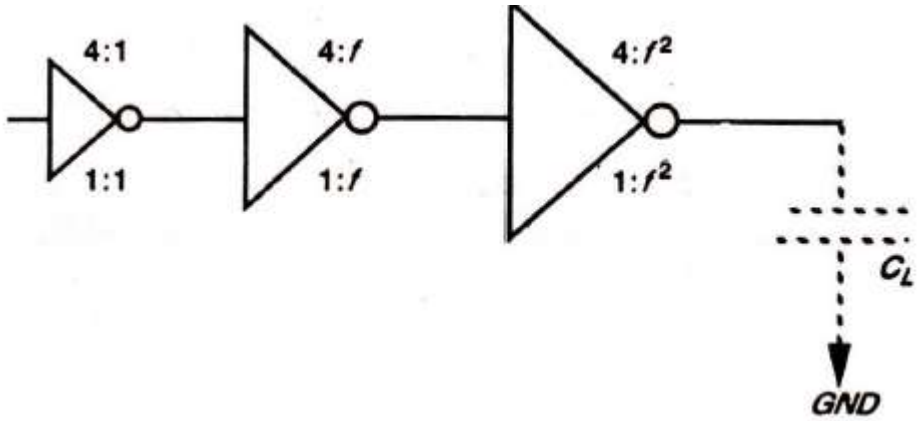


FIGURE Driving large capacitive loads

Inverters intended to drive large capacitive loads must therefore present low pull-up and pull-down resistance

Obviously, for MOS circuits, low resistance values for *Zp.d.* and *Zp.u.* imply low *L: W* ratios; in other words, channels must be made very wide to reduce resistance value and, in consequence, an inverter to meet this need occupies a large area. Moreover, because of the large *L: W* ratio and since length *L* cannot be reduced below the minimum feature size, the gate region area L x W becomes significant and a comparatively large capacitance is presented at the input, which in tum slows down the rates *of* change of voltage which can take place at the input.

The remedy is to use *N* cascaded inverters, each one of which is larger than the preceding stage by a width factor *f* as shown in Figure

Clearly, as the width factor increases, so the capacitive load presented at the inverter input increases, and the area occupied increases also. Equally clearly, the rate at which the width increases (that is, the value of.f) will influence the number N of stages which must be
cascaded to drive a particular value of C_L. Thus, an optimum solution must be sought as follows (

With large f, N decreases but delay per stage increases. For 4:1 nMOS inverters

$$\left.\begin{array}{r}\text{delay per stage} = f\tau \text{ for } \Delta V_{in} \\ \text{or} = 4f\tau \text{ for } \nabla V_{in}\end{array}\right\} \quad \begin{array}{l}\text{where } \Delta V_{in} \text{ indicates logic 0 to 1} \\ \text{transition and } \nabla V_{in} \text{ indicates} \\ \text{logic 1 to 0 transition of } V_{in}\end{array}$$

Therefore, total delay per nMOS pair = $5f\tau$. A similar treatment yields delay per CMOS pair = $7f\tau$. Now let

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

so that the choice of f and N are interdependent.

We now need to determine the value of f which will minimize the overall delay for a given value of y and from the definition of y

$$\ln(y) = N \ln(f)$$

That is

$$N = \frac{\ln(y)}{\ln(f)}$$

Thus, for N even

$$\text{total delay} = \frac{N}{2} 5f\tau = 2.5\ Nf\tau \text{ (nMOS)}$$

$$\text{or} = \frac{N}{2} 7f\tau = 3.5\ Nf\tau \text{ (CMOS)}$$

Thus, in all cases

$$\text{delay} \propto Nf\tau = \frac{\ln(y)}{\ln(f)} f\tau$$

It can be shown that total delay is minimized if f assumes the value e (base of natural logarithms); that is, each stage should be approximately 2.7* times wider than its predecessor. This applies to CMOS as well as nMOS inverters.

Thus, assuming that $f = e$, we have

$$\text{Number of stages } N = \ln(y)$$

and overall delay t_d

$$N \text{ even: } t_d = 2.5eN\,\tau \text{ (nMOS)}$$

$$\text{or } t_d = 3.5eN\,\tau \text{ (CMOS)}$$

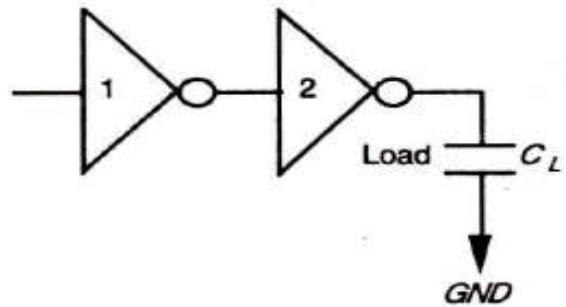
$$\left.\begin{aligned} N \text{ odd: } t_d &= [2.5(N-1)+1]e\tau \text{ (nMOS)} \\ \text{or } t_d &= [3.5(N-1)+2]e\tau \text{ (CMOS)} \end{aligned}\right\} \text{ for } \Delta V_{in}$$

or

$$\left.\begin{aligned} t_d &= [2.5(N-1)+4]e\tau \text{ (nMOS)} \\ \text{or } t_d &= [3.5(N-1)+5]e\tau \text{ (CMOS)} \end{aligned}\right\} \text{ for } \nabla V_{in}$$

Two nMOS inverters are cascaded to drive a capacitive load $C_L = 16 \square C_g$ as shown in Figure 4.20. Calculate the pair delay (V_{in} to V_{out}) in terms of τ for the inverter geometry indicated in the figure. What are the ratios of each inverter?

Assume a suitable value for τ and evaluate this pair delay.



Inverter 1	*Inverter 2*
$L_{p.u.} = 16\lambda$	$L_{p.u.} = 2\lambda$
$W_{p.u.} = 2\lambda$	$W_{p.u.} = 2\lambda$
$L_{p.d.} = 2\lambda$	$L_{p.d.} = 2\lambda$
$W_{p.d.} = 2\lambda$	$W_{p.d.} = 8\lambda$

Pair delay or total dealy ($t_{d)}$)=2.5Nfτ

Where N=2

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

$$\frac{16\square C_g}{\square C_g} = 16$$

$$f^N = 16 = 4^2$$

$$f = 4$$

($t_{d)}$)=2.5Nfτ

t_d=2.5X2X4Xτ=20τ

MOS LAYERS

MOS design is aimed at turning a specification into masks for processing silicon to meet the specification. We have seen that MOS circuits are formed on four basic *layers-n-diffusion*, *p-diffusiqn*, *polysilicon*, and *metal*, which are isolated from one another by thick or thin (thinox) silicon dioxide insulating layers

The thin oxide (thinox) mask region includes n-diffusion, p-diffusion, and transistor channels

.

Polysilicon and thinox regions interact so
that a transistor is formed where they cross one another. In some processes, there may be a
second metal layer and also, in some processes, a second polysilicon layer. Layers may deliberatelY joined together where contacts are formed.

STICK DIAGRAMS

Stick diagrams may be used to convey layer information through the use of a color codefor example, in the case of nMOS design, green for n-diffusion, red for polysilicon, blue for metal, yellow for implant, and black for contact areas.

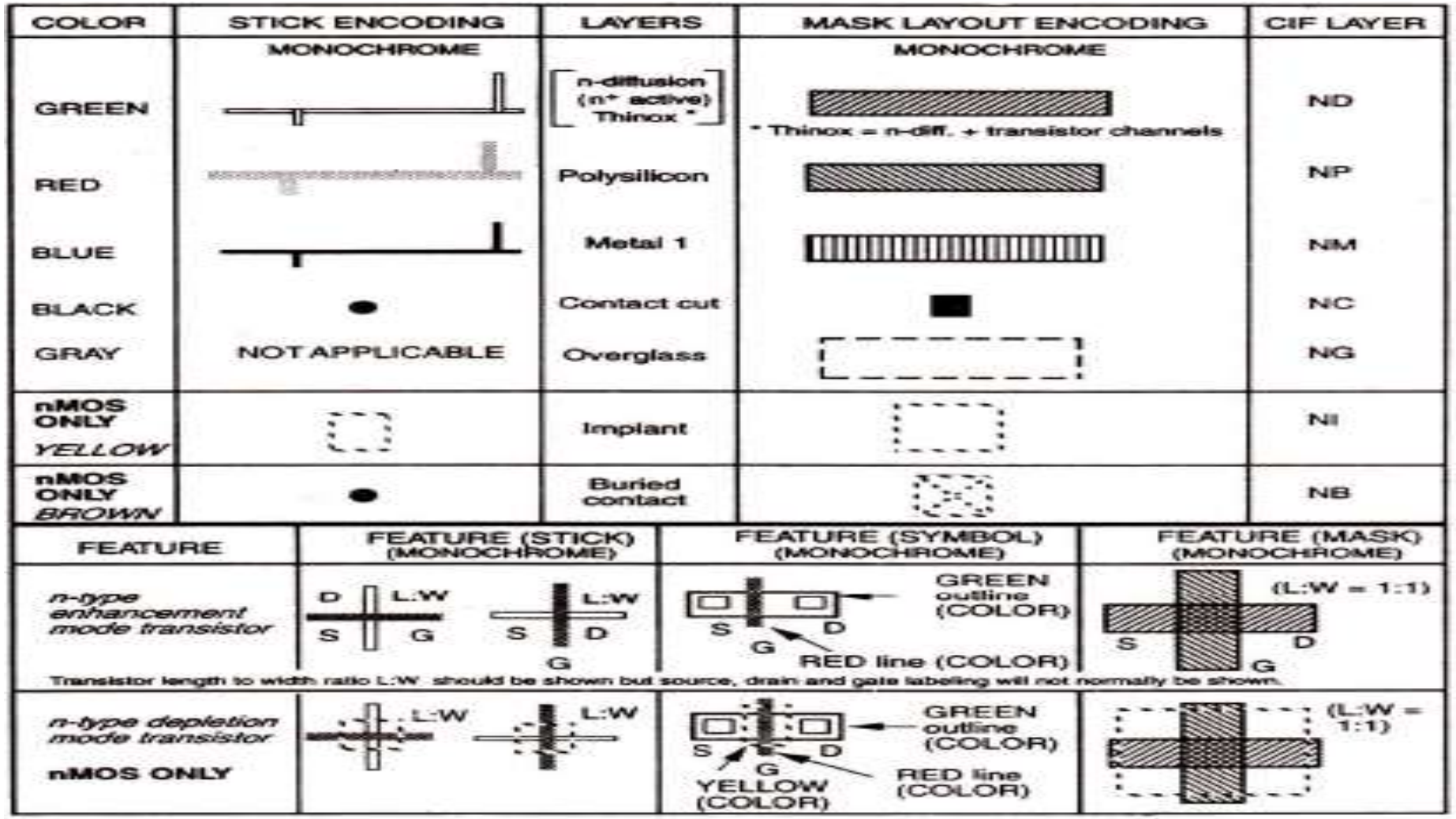
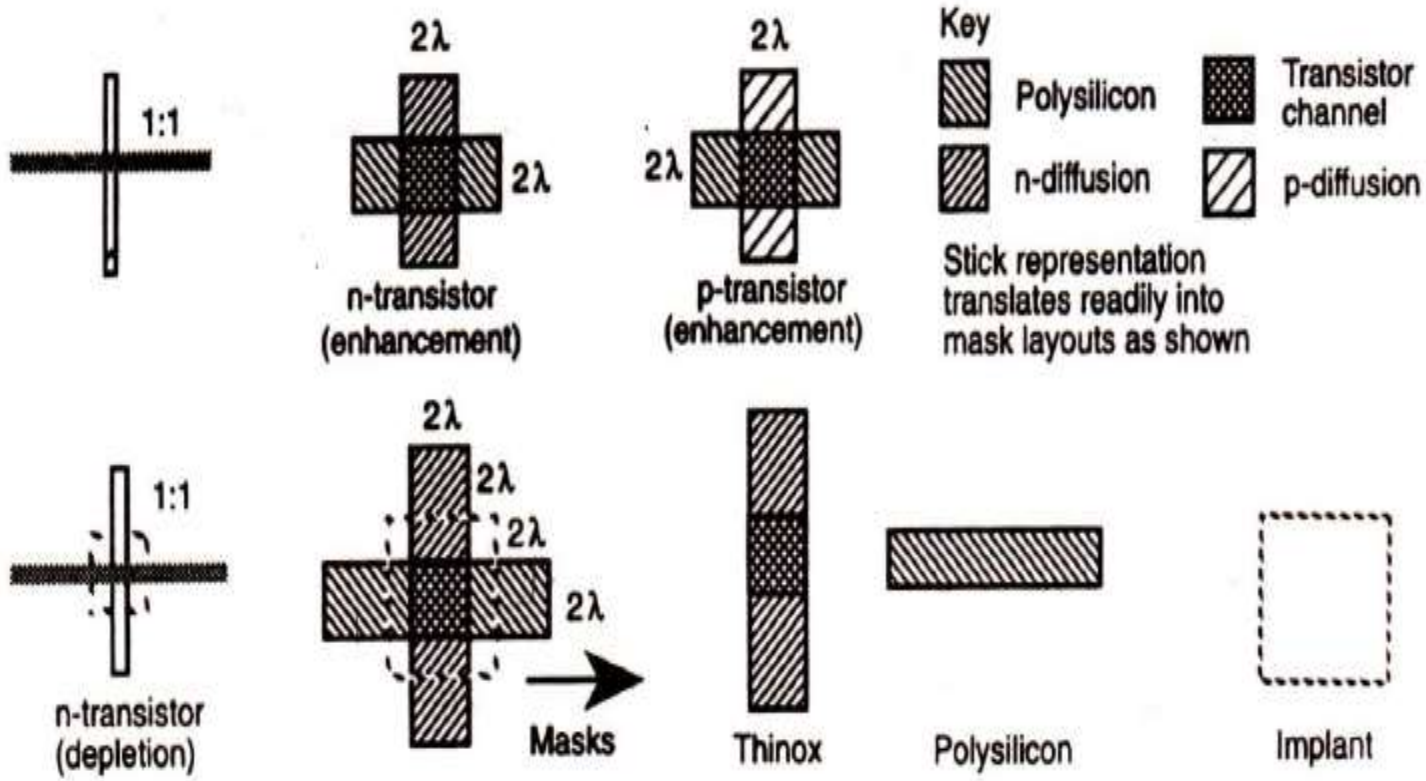


FIGURE :Encodings for a simple metal nMOS process

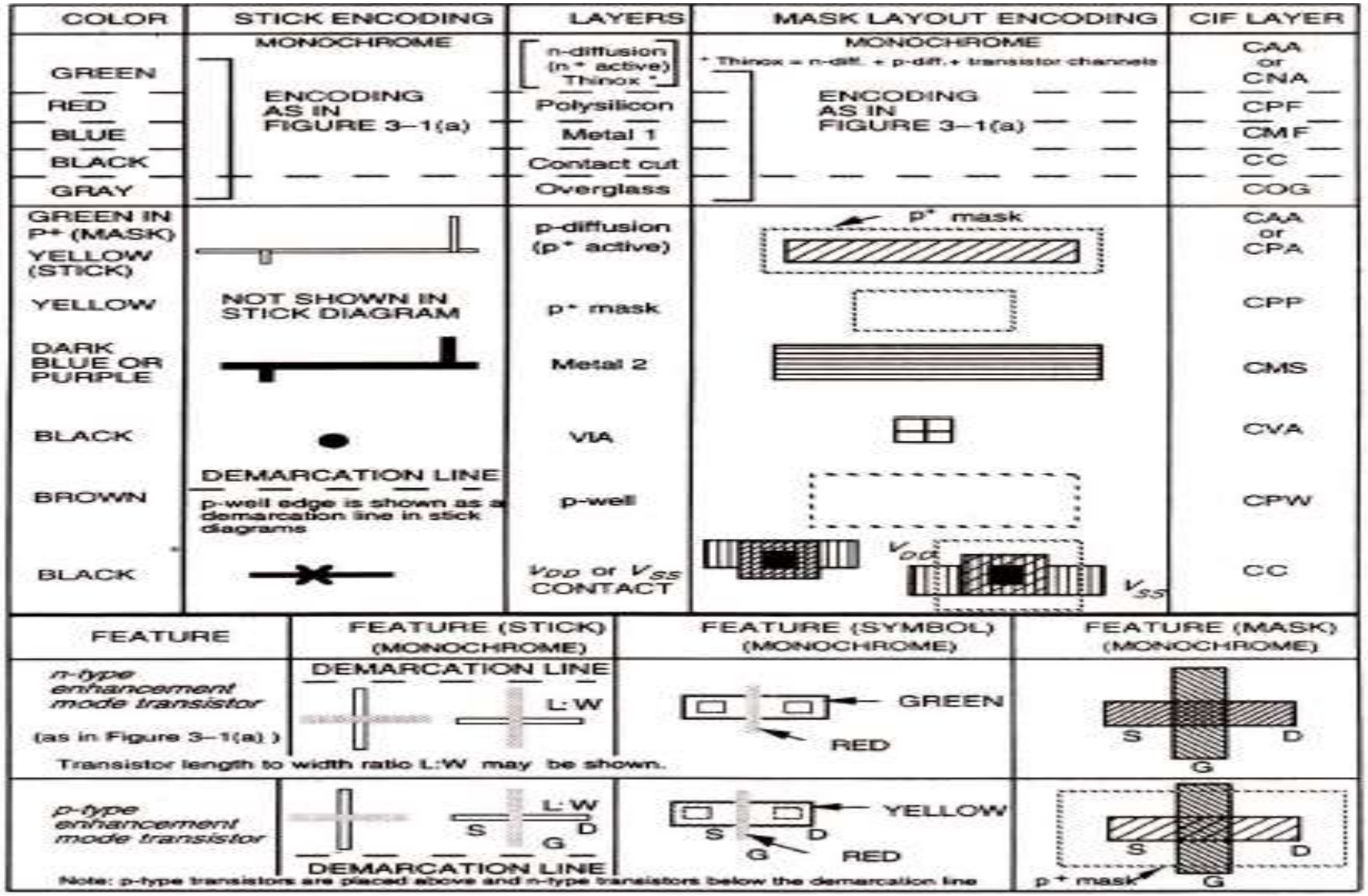

1:1
2λ
2λ
n-transistor
(enhancement)
2λ
2λ
p-transistor
(enhancement)
Key
Polysilicon
Transistor
channel
n-diffusion
p-diffusion
Stick representation
translates readily into
mask layouts as shown
1:1
n-transistor
(depletion)
2λ
2λ
2λ
2λ
Masks
Thinox
Polysilicon
Implant

STICK DIAGRAMS

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

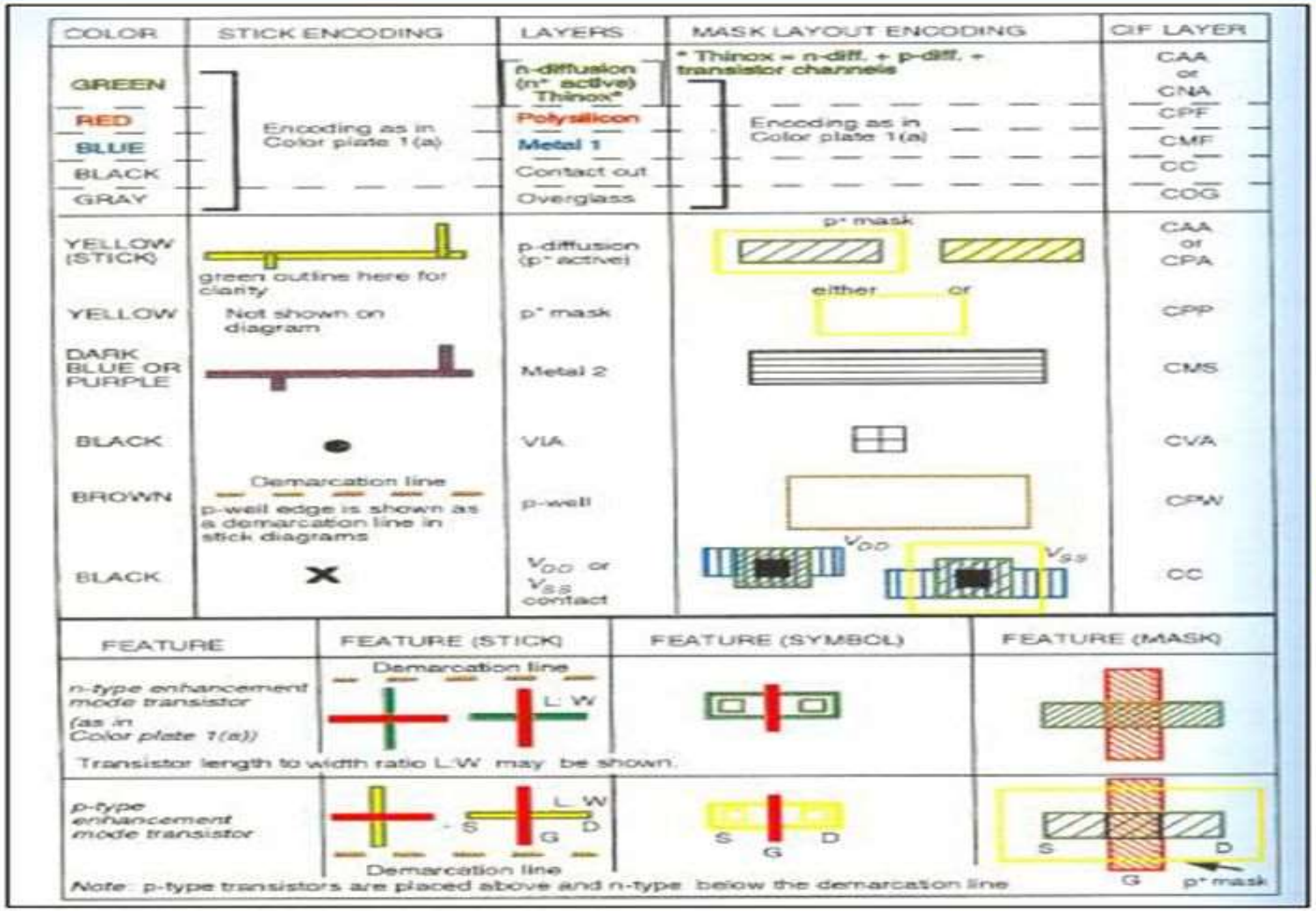
FEATURE	FEATURE (STICK)	FEATURE (SYMBOL)	FEATURE (MASK)
n-type enhancement mode transistor	D, L/W, S, G	L/W, S, D, G	(L/W = 1:1), S, D, G
Transistor length to width ratio L/W should be shown.			
n-type depletion mode transistor nMOS only	L/W	L/W, S, G, D	(L/W = 1:1)
Source, drain and gate labelling will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

NMOS ENCODING



Encodings for a double metal CMOS p-well process

STICK DIAGRAMS



CMOS ENCODING

nMOS Design Style

In order to start with a relatively simple process, we will consider single metal, single polysilicon nMOS technology

A rational approach to stick diagram layout is readily adopted for such nMOS circuits and the approach recommended here is both easy to use and to tum into a mask layout. The layout of nMOS involves:

- n-diffusion [n-diff.] and other thinoxide regions [thinox] (green);
- polysilicon I [poly.]-since there is only one polysilicon layer here (red);
- metal I [metal]-since we use only one metal layer here (blue);
- implant (yellow);
- contacts (black or brown [buried]).

A transistor is formed wherever poly. crosses n-diff. (red over green) and all diffusion wires (interconnections) are n-type (green).

When starting a layout, the first step normally taken is to draw the metal (blue) *Vdd* and *GND* rails in parallel allowing enough space between them for the other circuit elements which will be required.
Next, thinox (green) paths may be drawn between the rails for inverters and inverter-based logic as shown in Figure , not forgetting to make contacts as appropriate.
Inverters and inverter-based logic comprise a pull-up structure, usually a depletion mode transistor, connected from the output point to *Vdd* and a pull-down structure of enhancement mode transistors suitably interconnected between the output point and *GND.*
This step in the process is illustrated in Figure , remembering that poly. (red) crosses thinox (green) wherever transistors are required.
Do not forget the implants (yellow) for depletion mode transistors and do not forget to write in the length to width (L: *W)* ratio for each transistor

Ratios are important, particularly in nMOS and nMOS-Iike circuits.

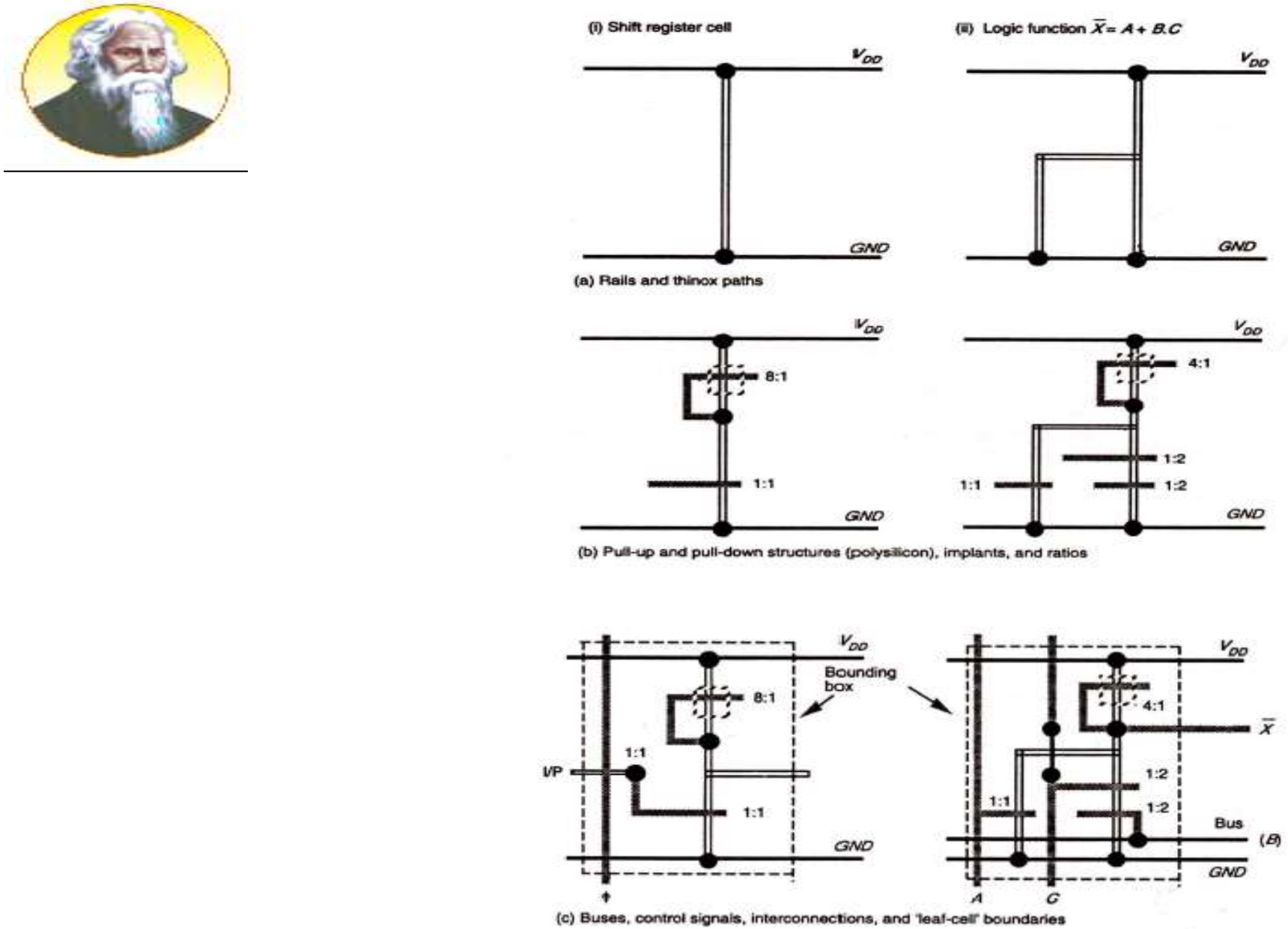


FIGURE: Examples of nMOS stick layout design style

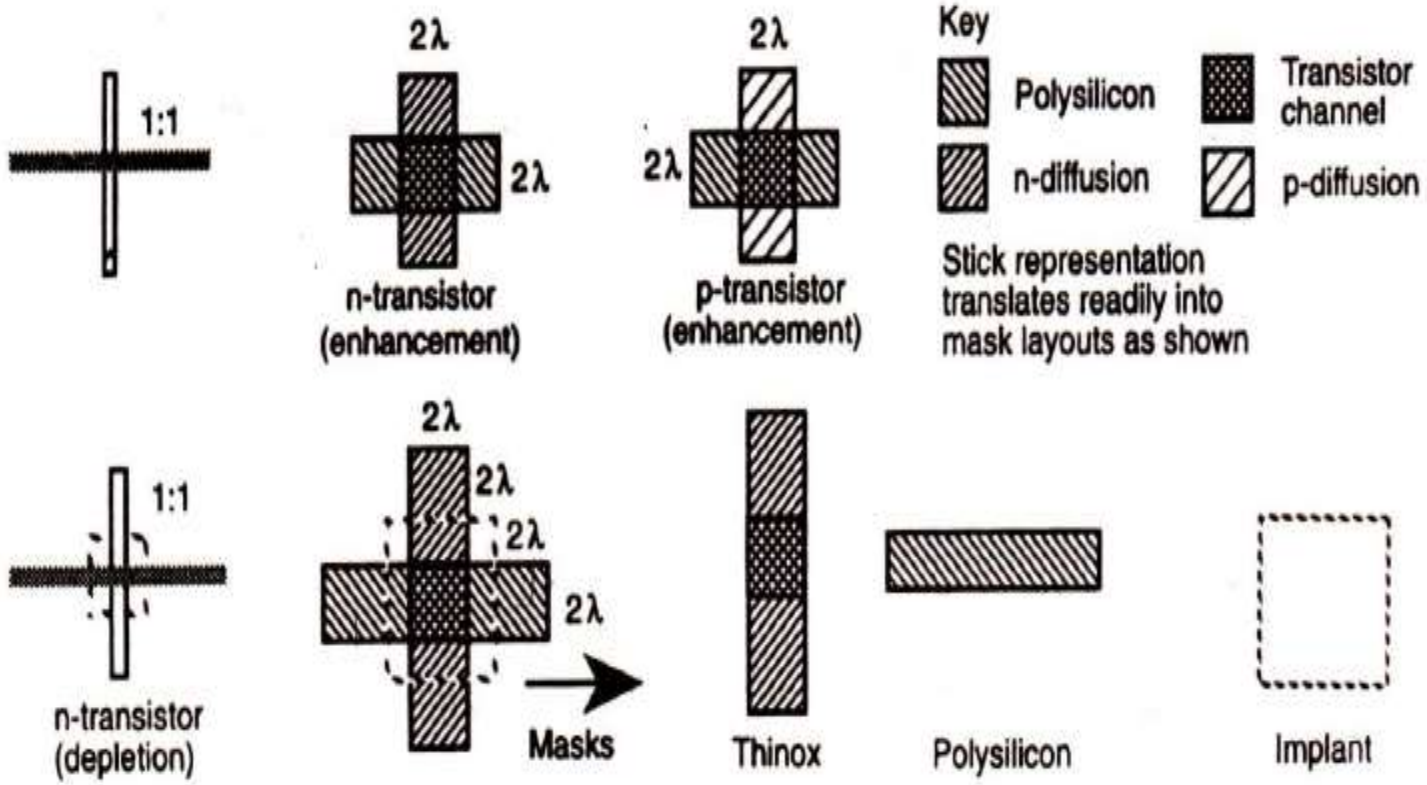
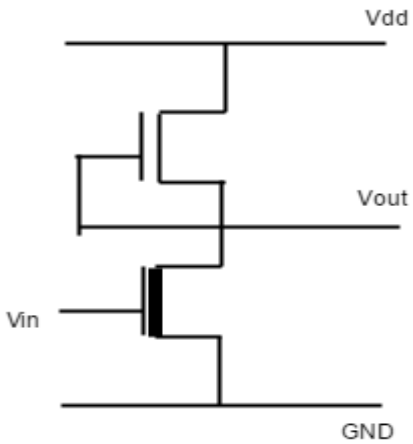
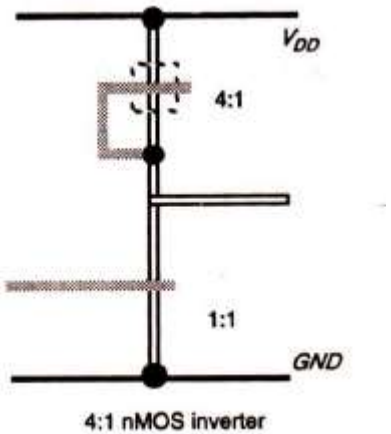
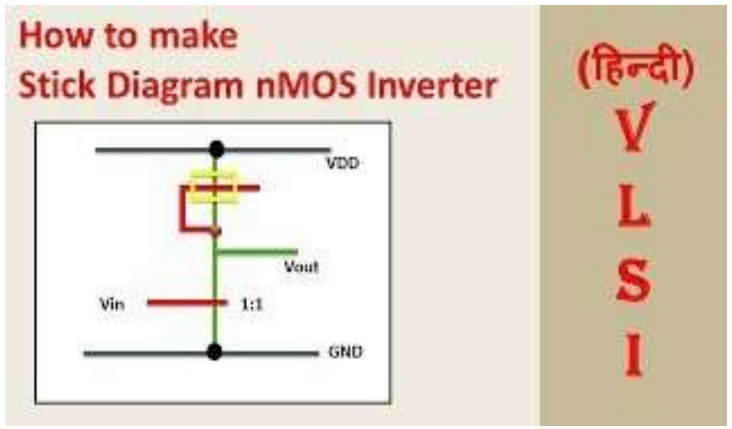


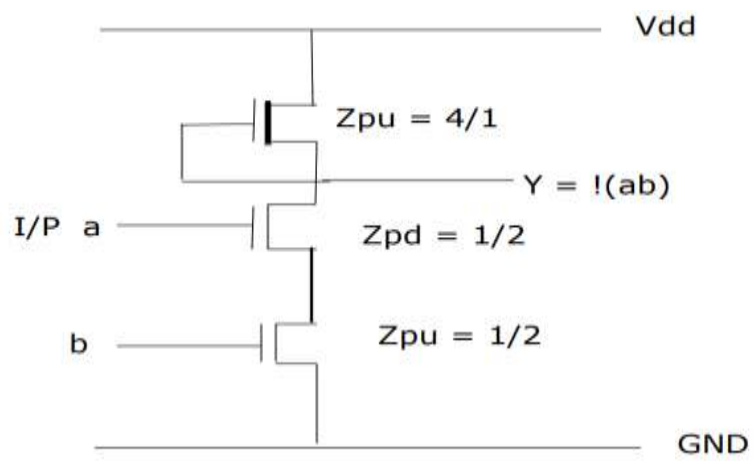
FIGURE Stick diagrams and corresponding mask layout examples.

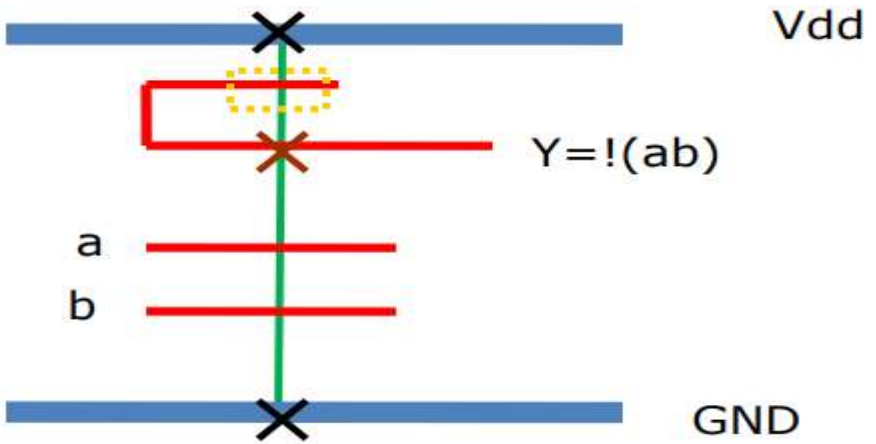




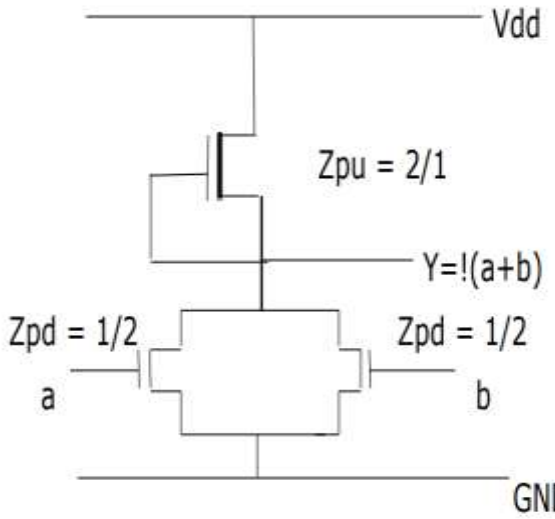
4:1 nMOS inverter

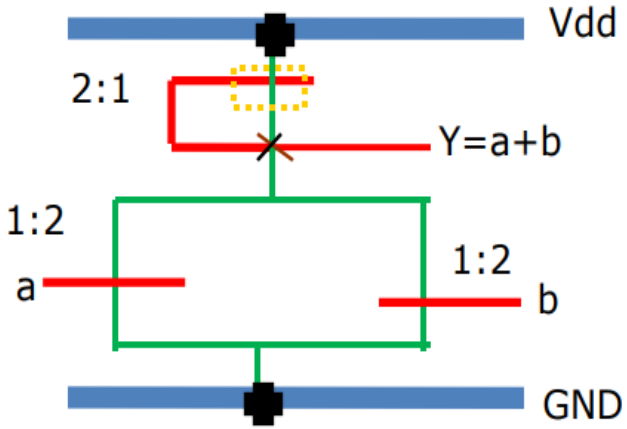






2 input NAND GATE USING NMOS





2 input NOR GATE USING NMOS

CMOS Design Style

All features and layers defined in Figure with the exception of implant (yellow) and the buried contact (brown), are used in CMOS design. Yellow in CMOS design is now used to identify p-transistors and wires, as depletion mode devices are not utilized. As a result, no confusion results from the allocation of the same color to two different features.

The two
types of transistor used, 'n' and 'p', are separated in the stick layout by the demarcation line (representing the p-well boundary) above which all p-type devices are placed (transistors and wires (yellow)). The n-devices (green) are consequently placed below the demarcation line and are thus located in the p-well.

FIGURE n-type and p-type transistors In CMOS design.

Diffusion paths must not cross the demarcation line and n-diffusion and p-diffusion wires must not join. The 'n' and 'p' features are normally joined by metal where a connection is needed.

However, we
must not forget to place crosses on *V DD* and *Vss* rails to represent the substrate and p-well
. connection respectively. The design begins with the drawing of the *Vdd* and
V ss rails in parallel and in metal and the creation of an (imaginary) demarcation line in between, as in Figure

The n-transistors are then placed below this line and thus close to Vss,while p transistors are placed above the line and below Vdd

It must be remembered that only metal and polysilicon can cross the demarcation line but with that restriction, wires can run in diffusion also

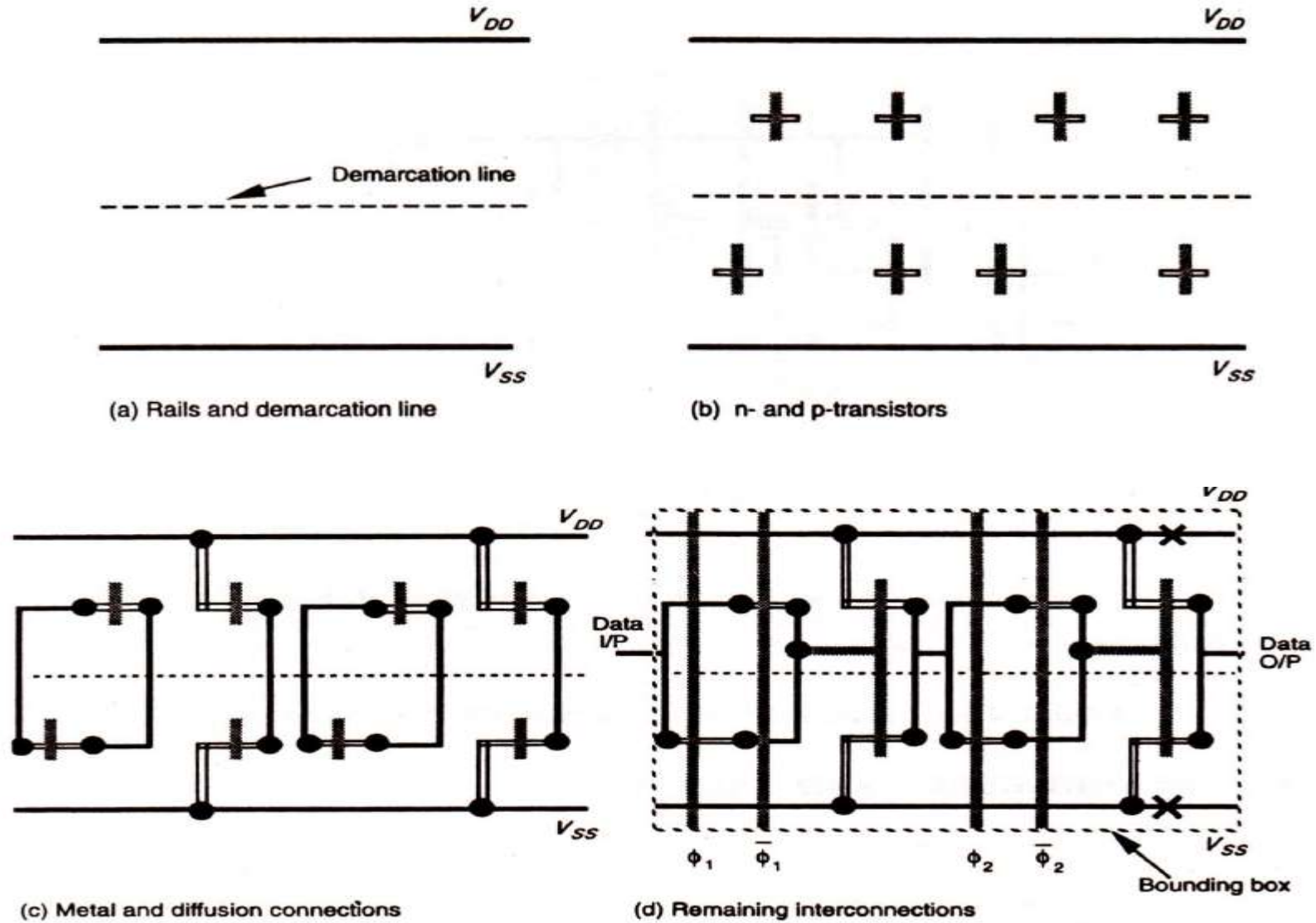
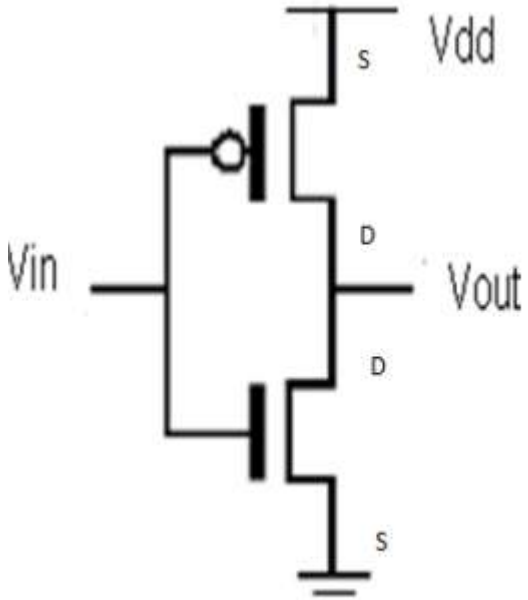
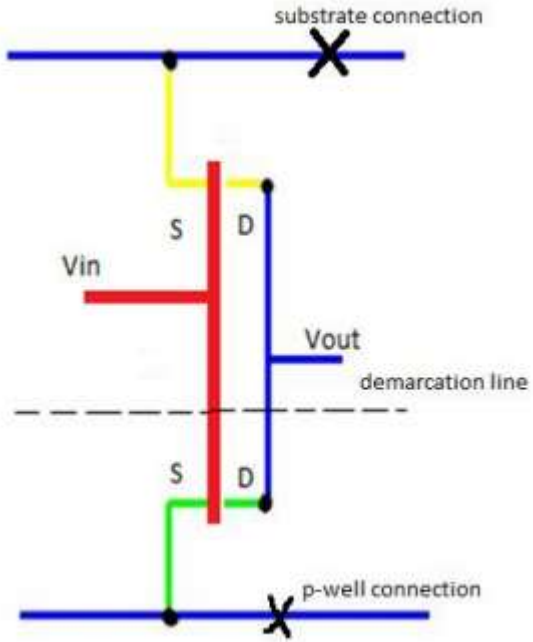
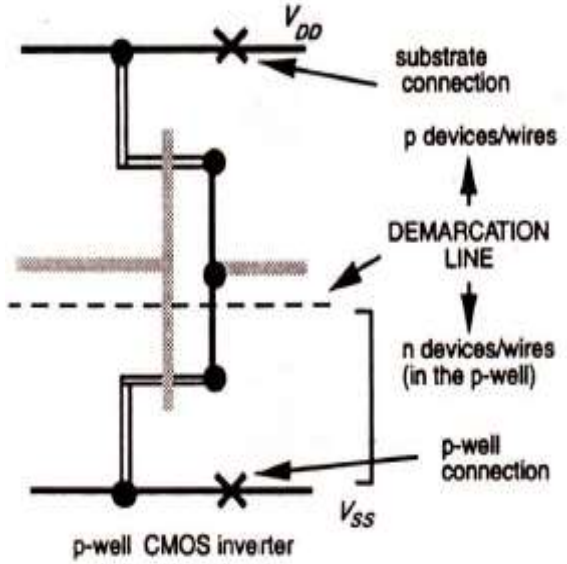


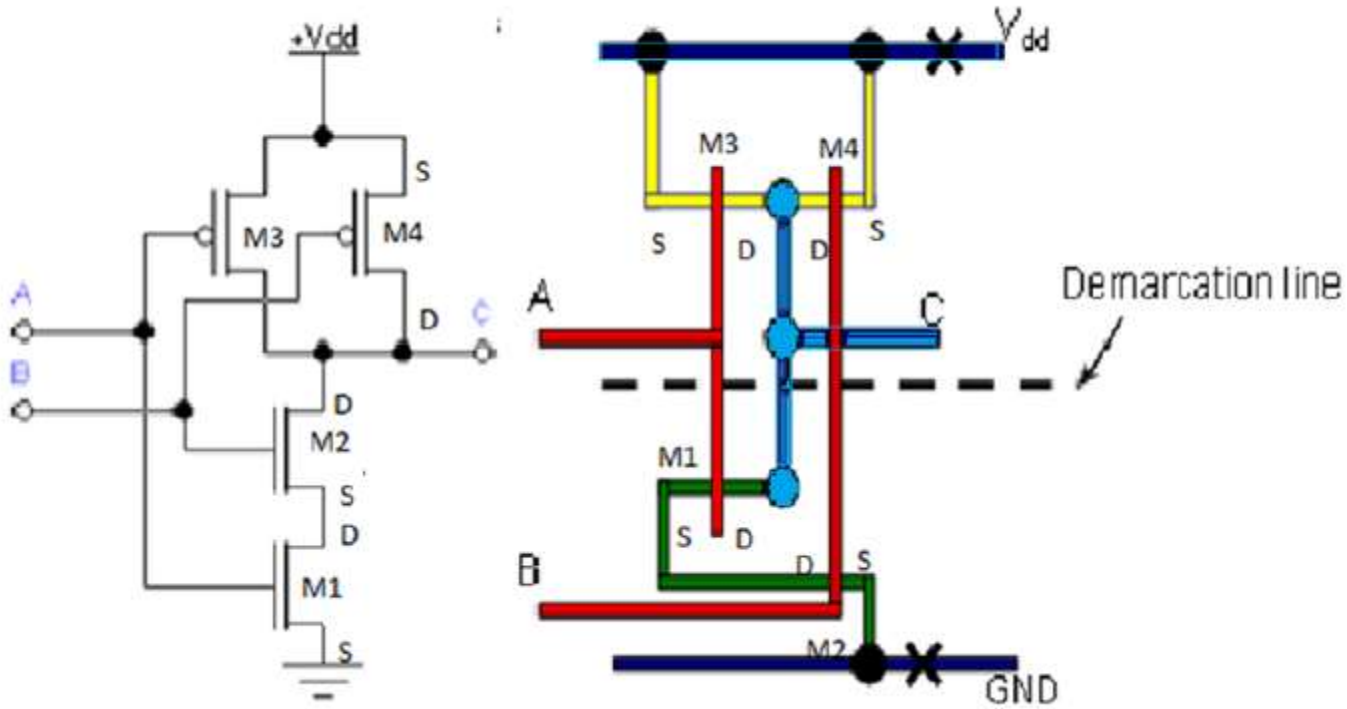
FIGURE Example of CMOS stick layout design style.



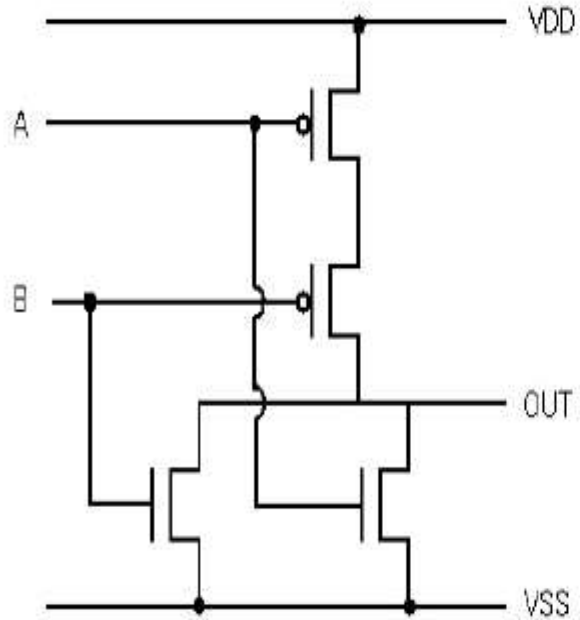




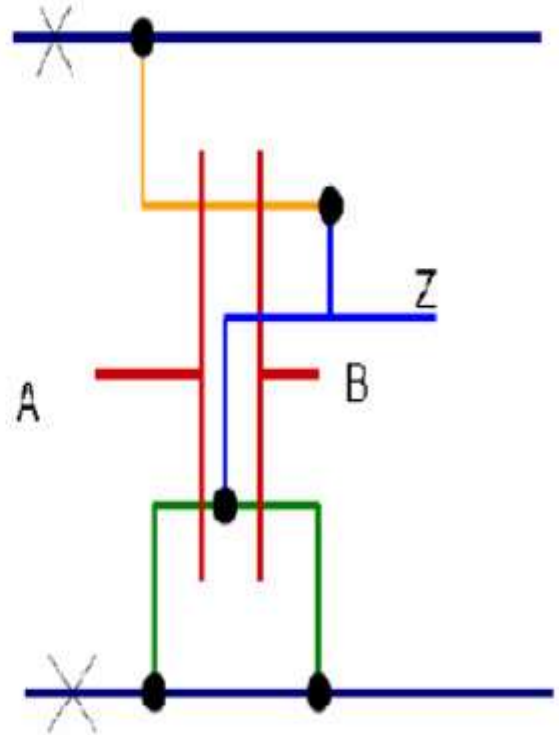
p-well CMOS inverter



2 INPUT –CMOS NAND GATE

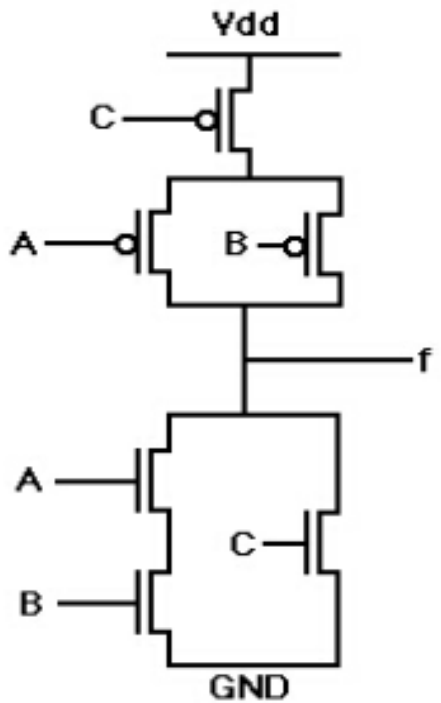


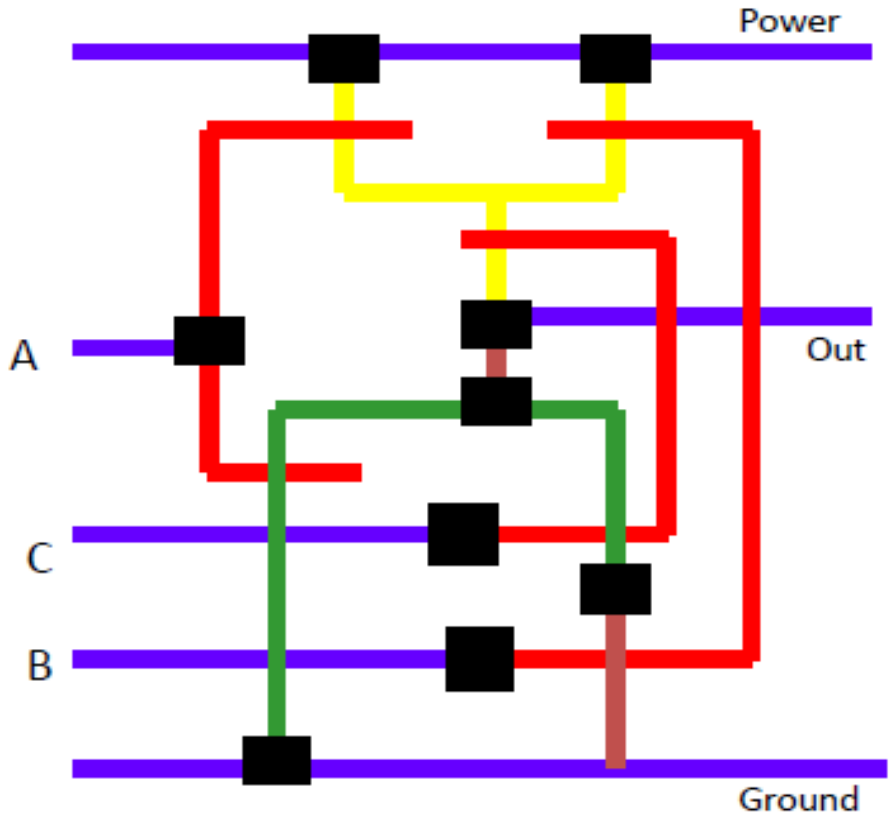
NOR gate in CMOS



2 INPUT –CMOS NOR GATE

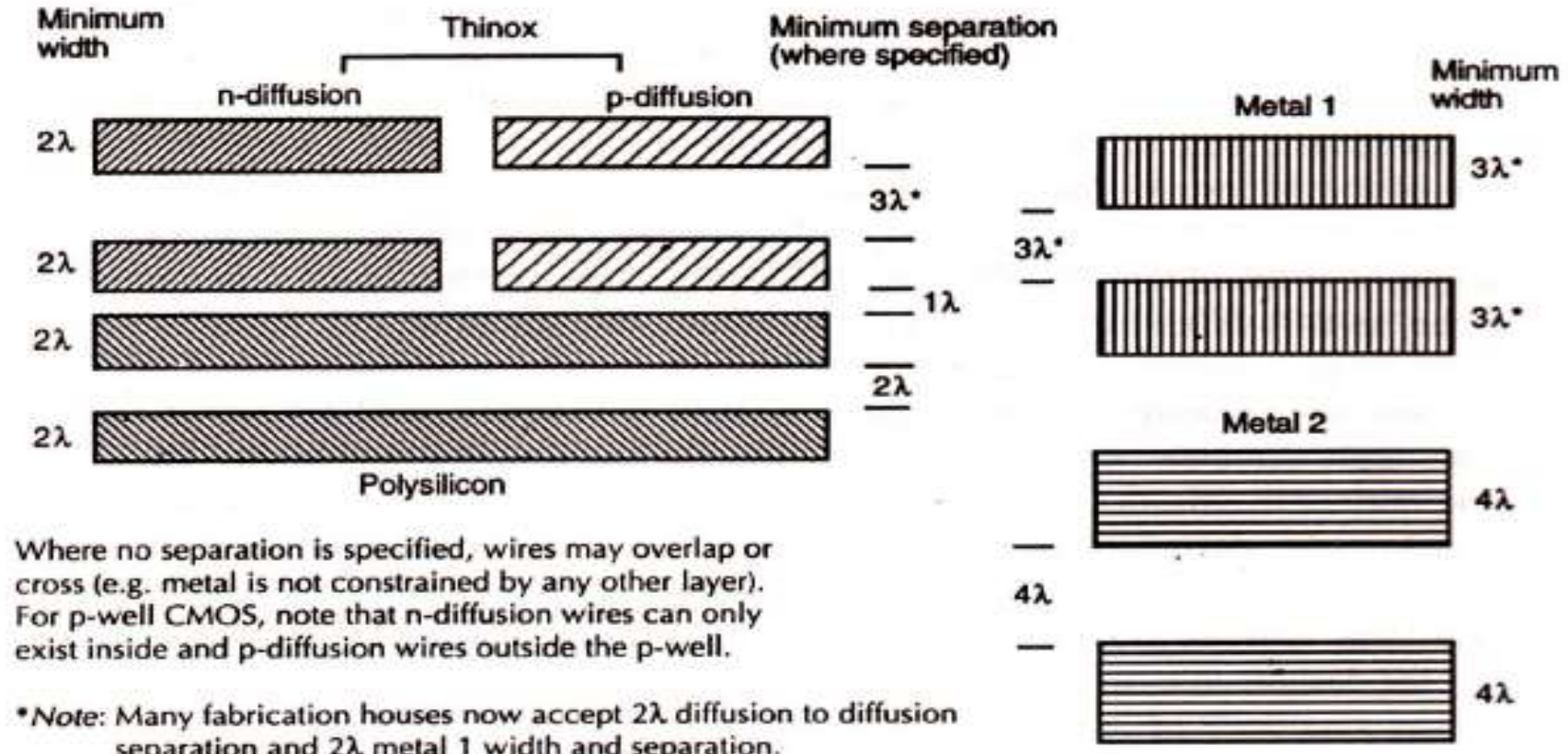
Example: $f = \overline{(A \bullet B) + C}$





DESIGN RULES AND LAYOUT

Lambda-based Design Rules



Design rules for wires (nMOS and CMOS)

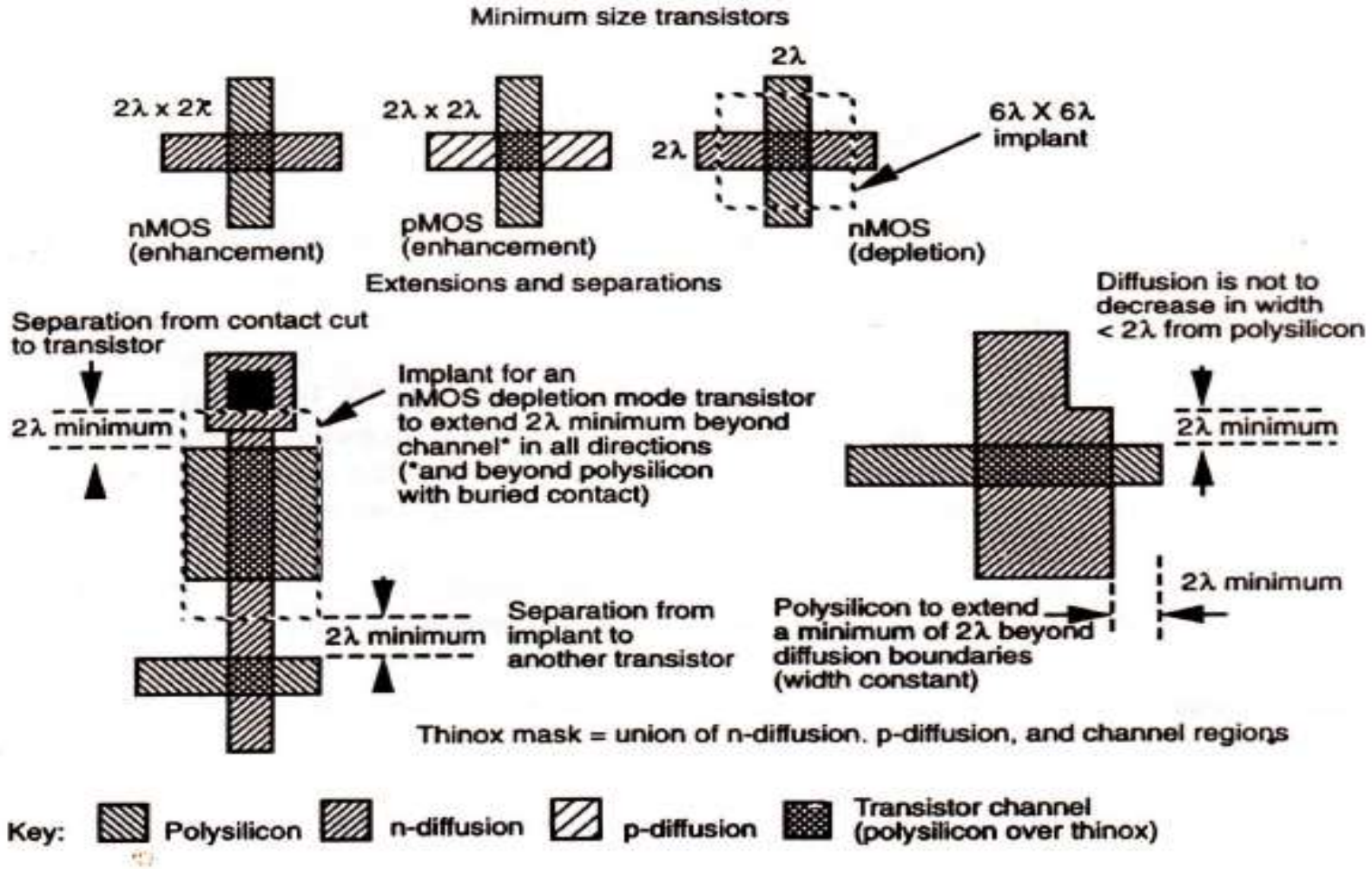


FIGURE Transistor design rules (nMOS, pMOS and CMOS).

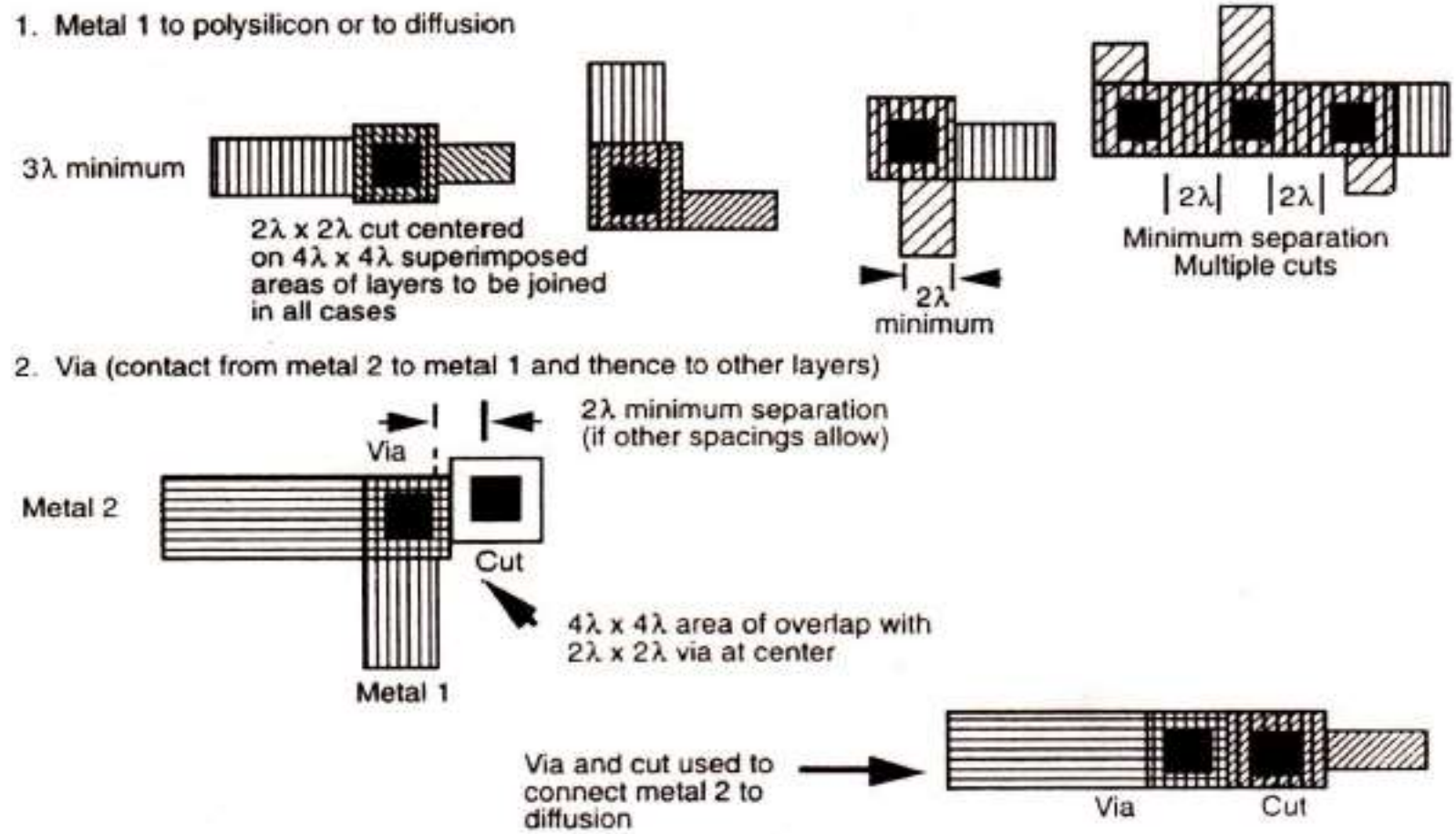


FIGURE Contacts (nMOS and CMOS).

2 µm DOUBLE METAL, DOUBLE POLY. CMOS/BICMOS RULES

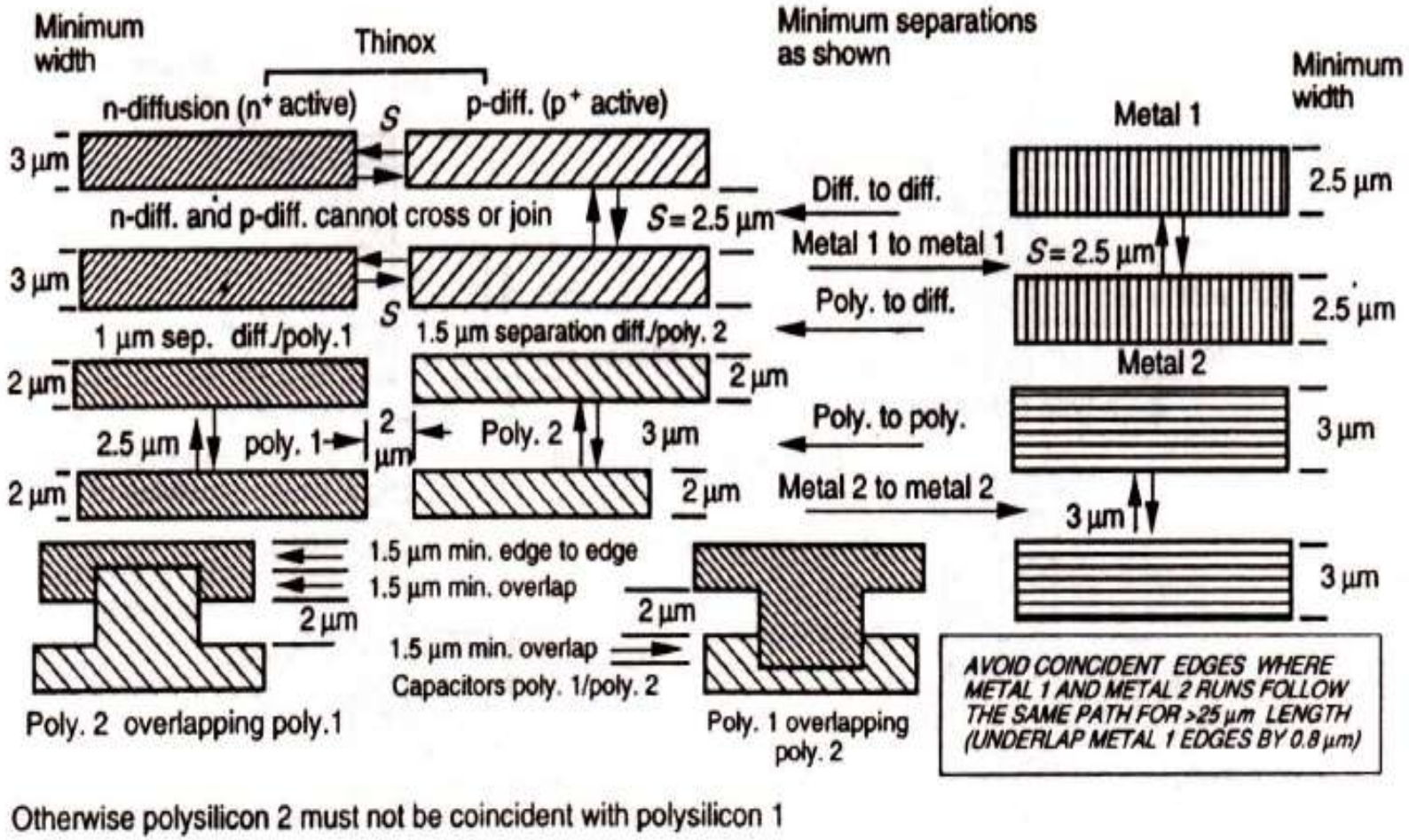
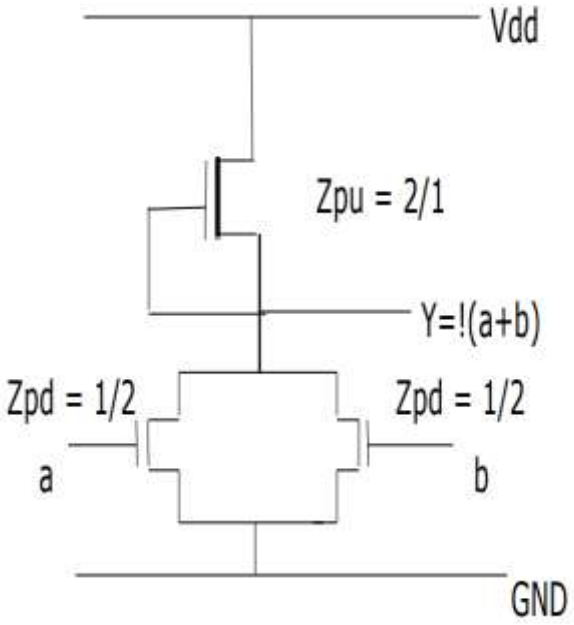
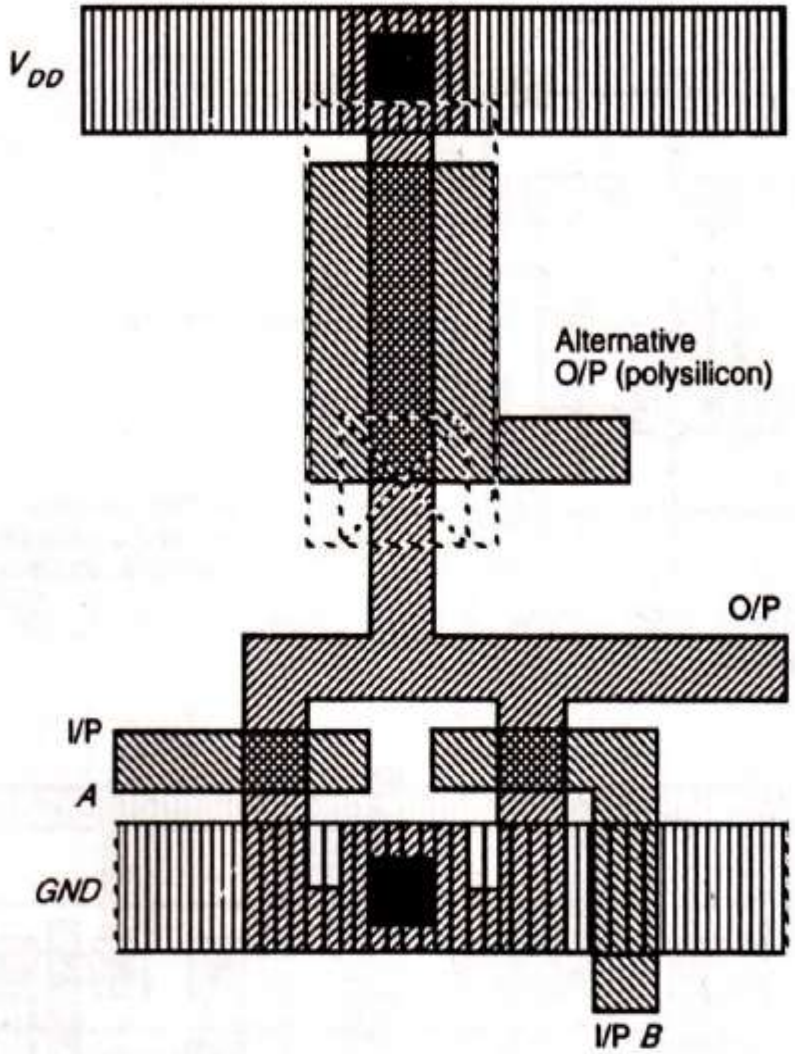


FIGURE Design rules for wires (interconnects) (Orbit 2 µm CMOS).





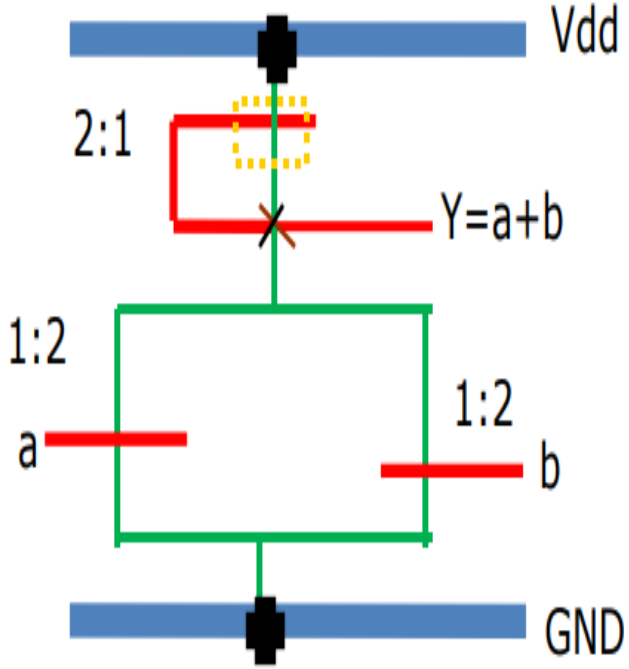
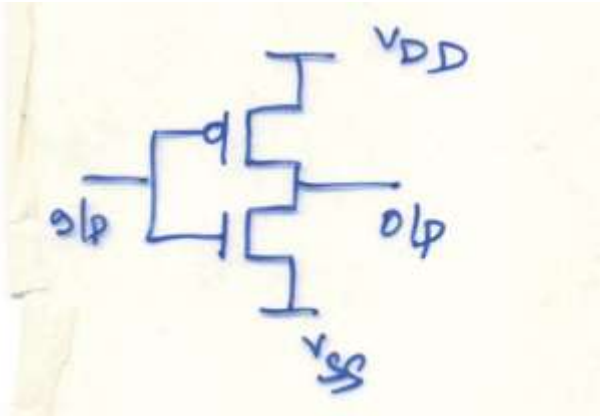
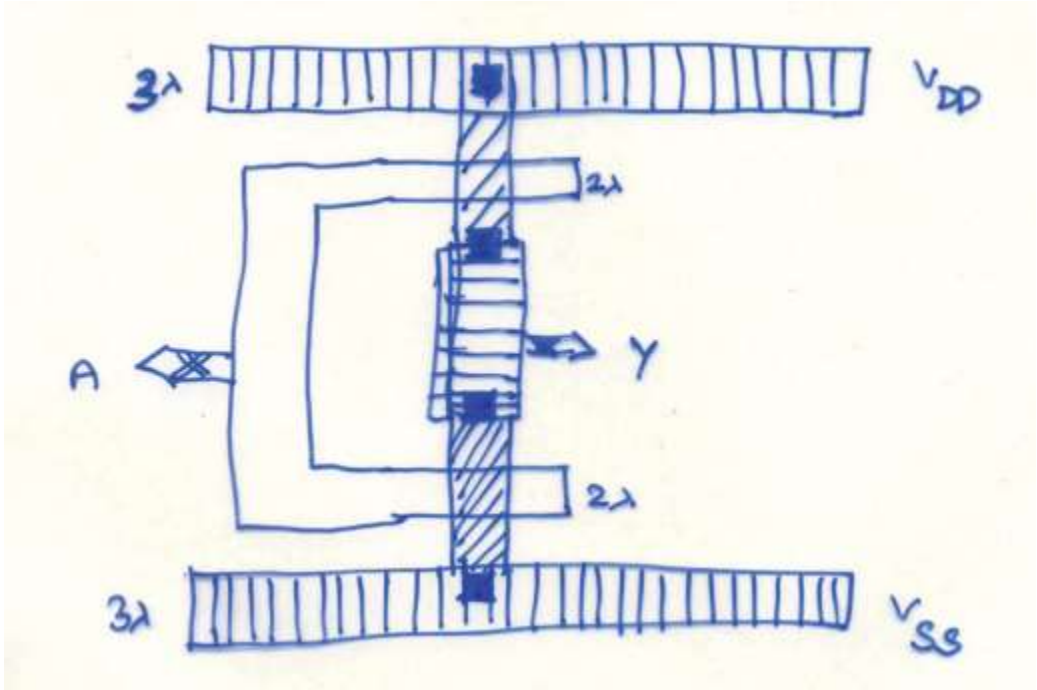


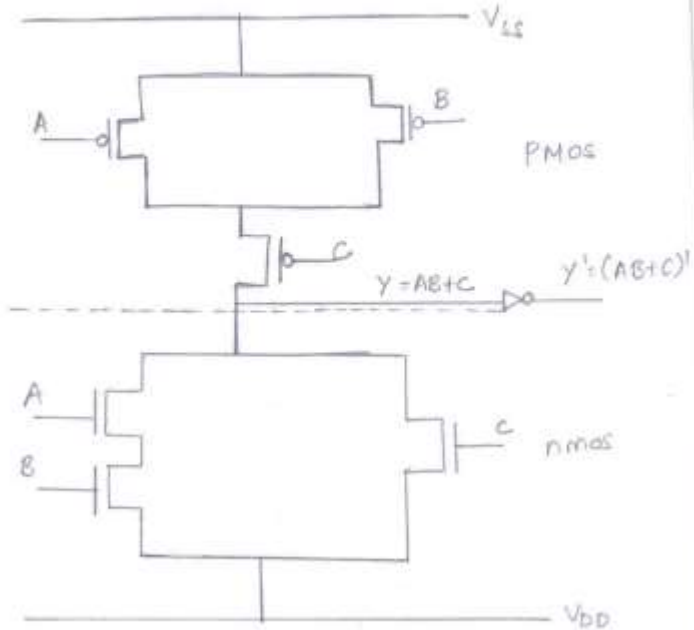
FIGURE Two 1/P n.MOS nor gate

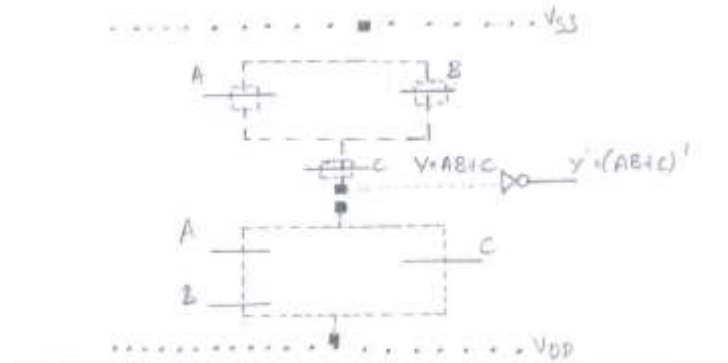

VDD
I/p
I/p
VSS

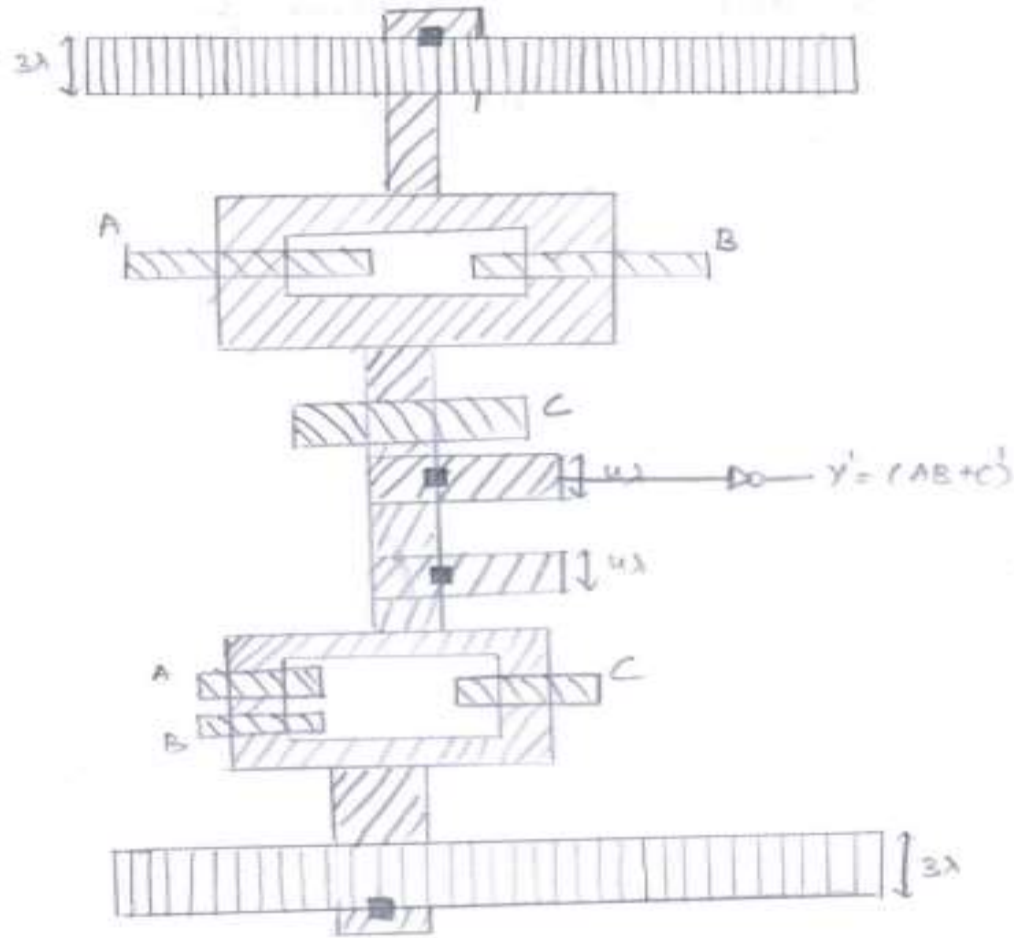

3λ
VDD
2λ
A
Y
2λ
3λ
VSS

Stick and layout diagram for

$Y = (AB + C)'$

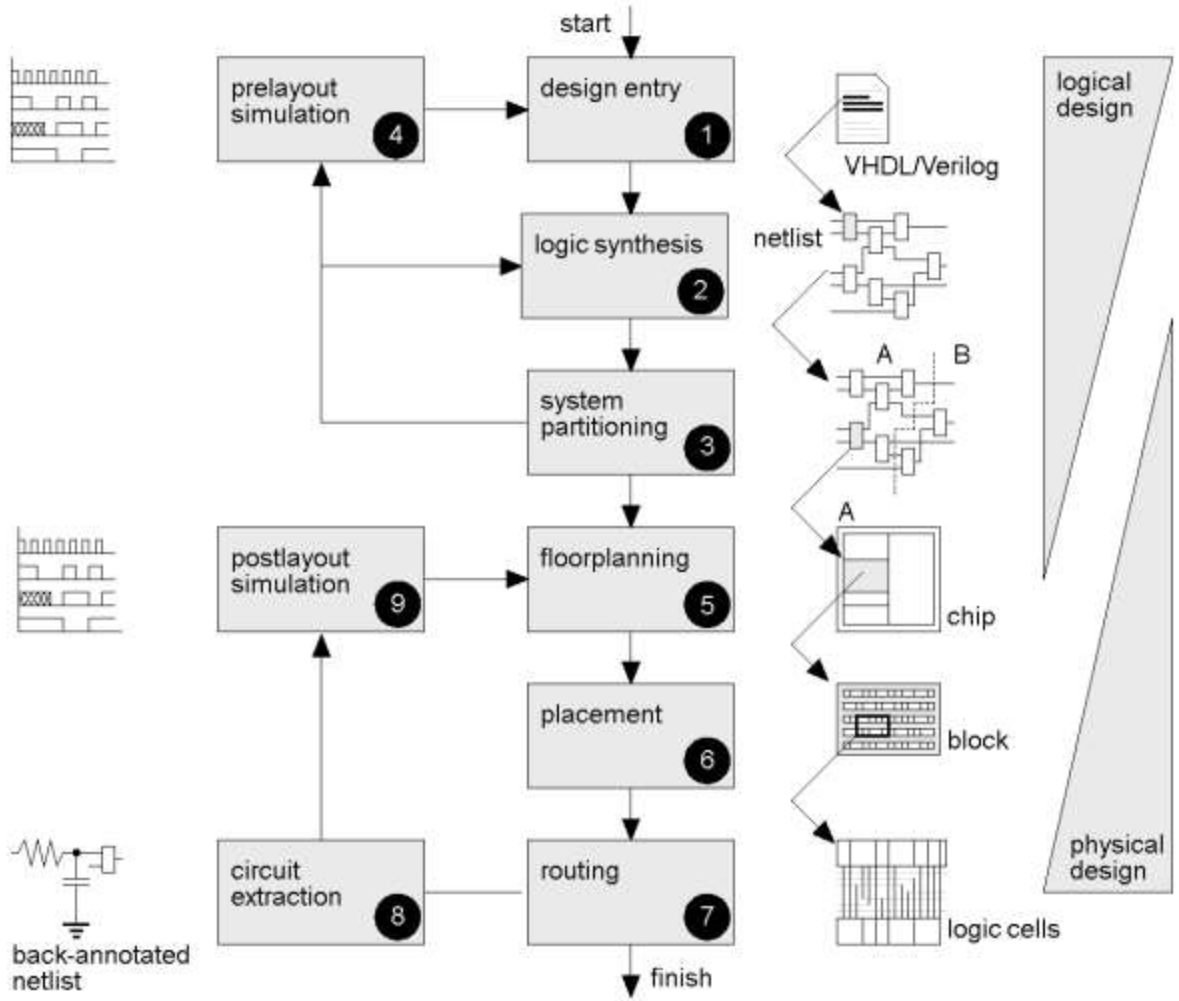





3λ
A
B
C
4λ
Y' = (AB+C)'
A
B
C
3λ

VLSI DESIGN FLOW

A design flow is a sequence of operations that transform the IC designers' intention (usually represented in RTL format) into layout GDSII data. A well-tuned design flow can help designers go through the chip-creation process relatively smoothly and with a decent chance of error-free implementation. And, a skilful IC implementation engineer can use the design flow creatively to shorten the design cycle, resulting in a higher likelihood that the product will catch the market window


start
prelayout simulation
4
design entry
1
logic synthesis
2
system partitioning
3
floorplanning
5
placement
6
routing
7
finish
circuit extraction
8
postlayout simulation
9
back-annotated netlist
VHDL/Verilog
netlist
A
B
A
chip
block
logic cells
logical design
physical design

Front-end design (Logical design):

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

2. Logic synthesis – Generation of net list (logic cells and their connections) from HDL code. Logic synthesis consists of following steps: (i) Technology independent Logic optimization (ii) Translation: Converting Behavioral description to structural domain (iii) Technology mapping or Library binding

3. System partitioning - Divide a large system into ASIC-sized pieces

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

Back-end design (Physical design):

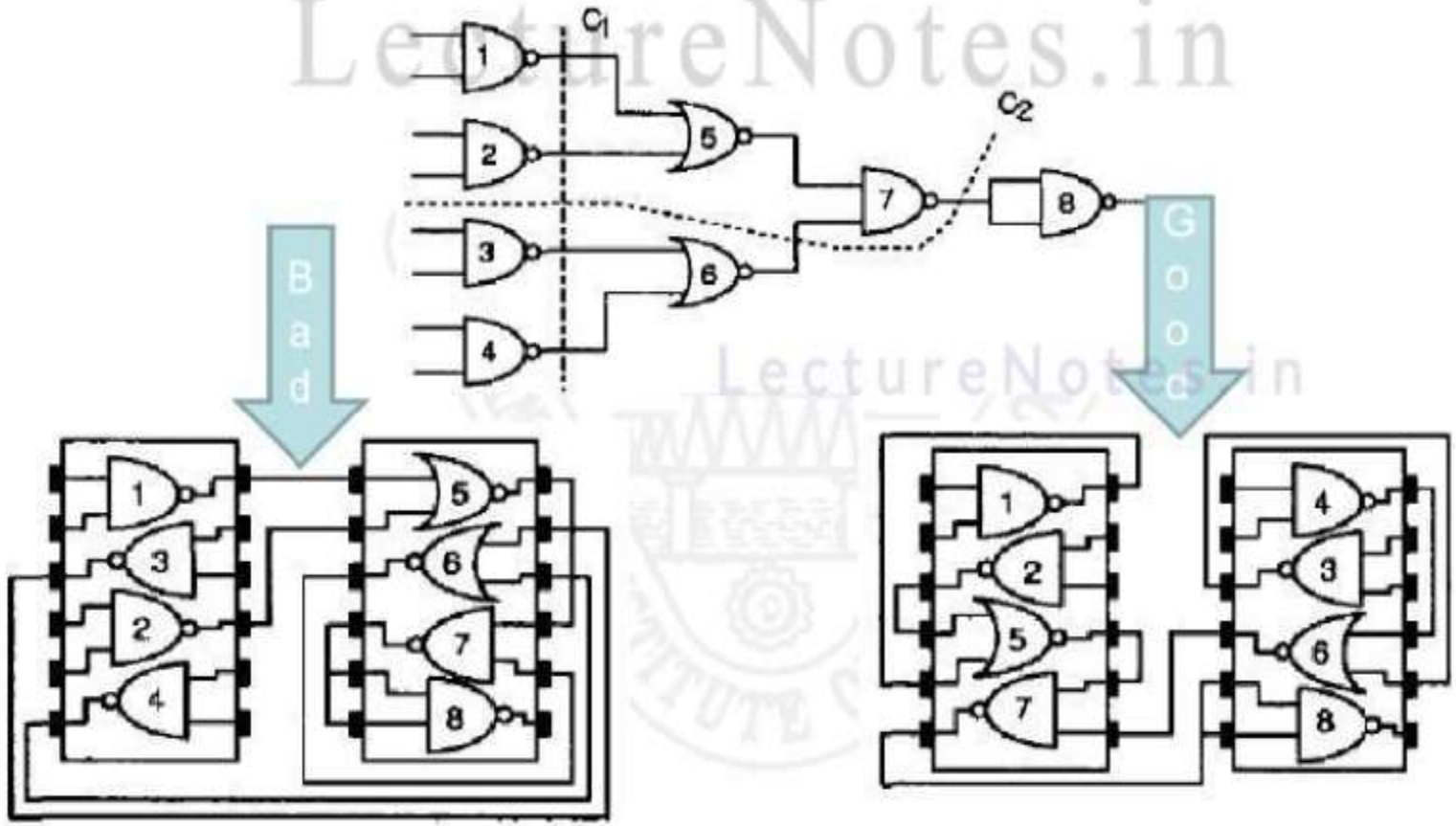
5. Floor planning - Arrange the blocks of the netlist on the chip

6. Placement - Decide the locations of cells in a block

7. Routing - Make the connections between cells and blocks

8. Circuit Extraction - Determine the resistance and capacitance of the interconnect

9. Post-layout simulation - Check to see the design still works with the added loads of the interconnect


C1
C2
1
2
3
4
5
6
7
8
Bad
Good

Scaling of MOS Circuits

Microelectronic technology may be characterized in terms of several indicators, or figures of merit. Commonly, the following are used:

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

Many of these figures of merit can be improved by shrinking the dimensions of transistors, interconnections and the separation between features, and by adjusting the doping levels and
supply voltages

Advantages of Scaling :

The reduction in lateral dimensions of the MOSFET and interconnects size is known as 'scaling' of the geometric dimensions of the MOSFET.

The advantages of Scaling are as follows,

(1) Improved current driving capability improves the device characteristics.

(2) Due to small geometries the capacitance reduces.

(3) Improved interconnect technology reduces the RC delay.

(4) The multiple threshold devices due to scaling adjusts the active and stand by power trade-offs.

(5) The integration density improves due to single chip devices.

(6) Enhanced performance in terms of speed and power consumption.

(7) Cost of a chip decreases by twice.

SCALING MODELS AND SCALIJ.'IIG FACTORS

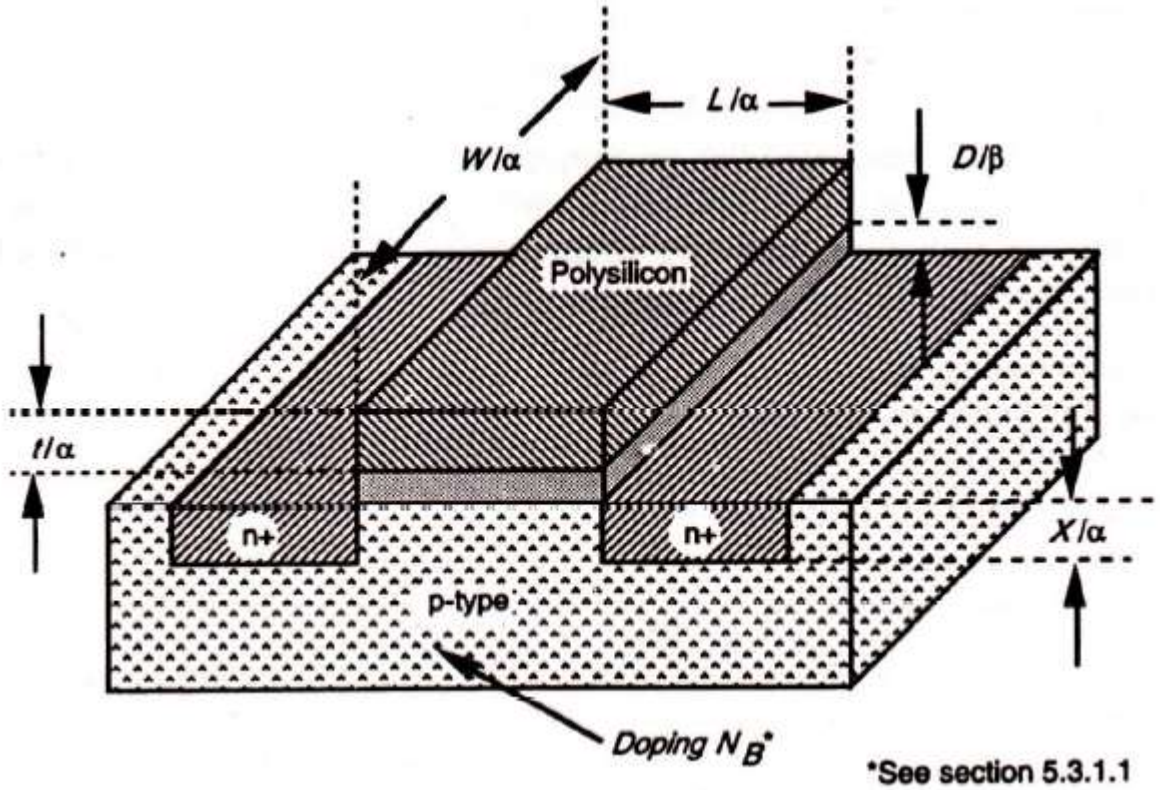


FIGURE Scaled nMOS transistor (pMOS similar).

SCALING MODELS AND SCALIJ.'IIG FACTORS

The most commonly used models are the constant electric field. scaling model and the constant
voltage scaling model. They both present a simplified view, taking only first degree effects into consideration, but are easily understood and well suited to educational needs. Recently, a combined voltage and dimension scaling model has been presented

In order to accommodate the three models, two scaling factors-1/α and 1/β---ar:e used. 1/ β is chosen as the scaling factor for supply voltage V_{DD} and gate oxide thickness D, and 1/α is used for all other linear. dimensions, both vertical and horizontal to the chip surface. For the constant field model and the constant voltage model, β = α and β = 1 respectively are applied

5.2.1 Gate Area A_g

$$A_g = L.W.$$

where L and W are the channel length and width respectively. Both are scaled by $1/\alpha$.

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

5.2.2 Gate Capacitance Per Unit Area C_0 or C_{ox}

$$C_0 = \frac{\varepsilon_{ox}}{D}$$

where ε_{ox} is the permittivity of the gate oxide (thinox) [$= \varepsilon_{ins}.\varepsilon_0$] and D is the gate oxide thickness which is scaled by $1/\beta$

Thus C_0 is scaled by $\frac{1}{1/\beta} = \beta$

5.2.3 Gate Capacitance C_g

$$C_g = C_0 L.W.$$

Thus C_g is scaled by $\beta \frac{1}{\alpha^2} = \frac{\beta}{\alpha^2}$

5.2.4 Parasitic Capacitance C_x

C_x is proportional to $\frac{A_x}{d}$

where d is the depletion width around source or drain which is scaled by $1/\alpha$, and A_x is the area of the depletion region around source or drain which is scaled by $1/\alpha^2$.

Thus C_x is scaled by $\frac{1}{\alpha^2} \cdot \frac{1}{1/\alpha} = \frac{1}{\alpha}$

5.2.5 Carrier Density in Channel Q_{on}

$$Q_{on} = C_0 \cdot V_{gs}$$

where Q_{on} is the average charge per unit area in the channel in the 'on' state. Note that C_0 is scaled by β and V_{gs} is scaled by $1/\beta$.

Thus Q_{on} is scaled by 1

5.2.6 Channel Resistance R_{on}

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

where μ is the carrier mobility in the channel and is assumed constant.

Thus R_{on} is scaled by $\frac{1}{\alpha} \frac{1}{1/\alpha} 1 = 1$

5.2.7 Gate Delay T_d

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

Thus T_d is scaled by $\frac{1.\beta}{\alpha^2} \frac{\beta}{\alpha^2}$

5.2.8 Maximum Operating Frequency f_0

$$f_0 = \frac{W}{L} \frac{\mu C_0 V_{DD}}{C_g}$$

or, f_0 is inversely proportional to delay T_d.

Thus f_0 is scaled by $\frac{1}{\beta/\alpha^2} = \frac{\alpha^2}{\beta}$

5.2.9 Saturation Current I_{dss}

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

noting that both V_{gs} and V_t are scaled by $1/\beta$, we have

I_{dss} is scaled by $\beta(1/\beta)^2 = 1/\beta$

5.2.10 Current Density *J*

$$J = \frac{I_{dss}}{A}$$

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

So, J is scaled by $\dfrac{1/\beta}{1/\alpha^2} = \dfrac{\alpha^2}{\beta}$

5.2.11 Switching Energy Per Gate E_g

$$E_g = \frac{1C_g}{2}(V_{DD})^2$$

So, E_g is scaled by $\dfrac{\beta}{\alpha^2} \cdot \dfrac{1}{\beta^2} = \dfrac{1}{\alpha^2\beta}$

5.2.12 Power Dissipation Per Gate P_g

P_g comprises two components such that

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

where the static component

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

and the dynamic component

$$P_{gd} = E_g f_0$$

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

So, P_g is scaled by $1/\beta^2$

5.2.13 Power Dissipation Per Unit Area P_a

$$P_a = \frac{P_g}{A_g}$$

So, P_a is scaled by $\dfrac{1/\beta^2}{1/\alpha^2} = \alpha^2/\beta^2$

5.2.14 Power-speed Product P_T

$$P_T = P_g \cdot T_d$$

So, P_T is scaled by $\dfrac{1}{\beta^2} \cdot \dfrac{\beta}{\alpha^2} = \dfrac{1}{\alpha^2\beta}$

Summary of Scaling Effects

Parameters		*Combined V and D*	*Constant E*	*Constant V*
V_{DD}	Supply voltage	$1/\beta$	$1/\alpha$	1
L	Channel length	$1/\alpha$	$1/\alpha$	$1/\alpha$
W	Channel width	$1/\alpha$	$1/\alpha$	$1/\alpha$
D	Gate oxide thickness	$1/\beta$	$1/\alpha$	1
A_g	Gate area	$1/\alpha^2$	$1/\alpha^2$	$1/\alpha^2$
C_0(or C_{ox})	Gate C per unit area	β	α	1
C_g	Gate capacitance	β/α^2	$1/\alpha$	$1/\alpha^2$
C_x	Parasitic capacitance	$1/\alpha$	$1/\alpha$	$1/\alpha$
Q_{on}	Carrier density	1	1	1
R_{on}	Channel resistance	1	1	1
I_{dss}	Saturation current	$1/\beta$	$1/\alpha$	1
A_c	Conductor X-section area	$1/\alpha^2$	$1/\alpha^2$	$1/\alpha^2$
I	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 dispn per gate	$1/\beta^2$	$1/\alpha^2$	1
N	Gates per unit area	α^2	α^2	α^2
P_a	Power dispn per unit area	α^2/β^2	1	α^2
T_d	Gate delay	β/α^2	$1/\alpha$	$1/\alpha^2$
f_0	Max. operating frequency	α^2/β	α	α^2
P_T	Power-speed product	$1/\alpha^2.\beta$	$1/\alpha^3$	$1/\alpha^2$

Constant E:$\beta = \alpha$; Constant V:$\beta = 1$

Limitations of Scaling

(1) Substrate doping. $d = \sqrt{\frac{2\epsilon_{si}\epsilon_0 V}{qN_B}}$ (where $V = V_a + V_B$)

(2) Limits on miniaturization

(3) limits of interconnect & contact resistance.

(4) Limits due to subthreshold currents.

(5) Limits on logic levels & supply Vtg due to noise.

(6) Limits due to current density.
[$J = 1$ to $2 mA/\mu m^2$]