UNIT – IV GATE LEVEL DESIGN

The gate of a MOS transistor controls the flow of current between the source and drain. Therefore the gate is a control input. Hence by using CMOS transistor all basic logic gates and complex logic gate can be implemented.

MOSFETs as Switches

Depending on the type of substrate used MOSFETs can be divided into two types : nMOS and pMOS. The gate terminal acts as a control input as it affects the electrical current between source and drain.

In nMOS transistor the body is generally grounded so that p-n junctions of the source and drain to body are reverse biased. Hence the transistor is said to OFF.

When the gate voltage is raised, it induces an electric field that attracts free electrons along Si-SiO_2 interface. When gate voltage raised to enough level a n-type channel is formed. The channel of electron carriers is formed from source to drain and transistor is said to ON.

In pMOS transistor, the situation is reversed, the body of transistor is held at high potential. The source and drain junctions are reverse biased and no current is flowing, so transistor is OFF. When gate voltage is lowered, positive charges are attracted to the underside of Si-SiO_2 interface. A sufficient low gate voltage inverts the channel and conducting path is formed between source and drain i.e. transistor is ON. The pMOS transistor symbol has a bubble on the gate showing the opposite behaviour of nMOS.

Briefly, MOS transistors can be used as ON/OFF switches. When the gate of an nMOS transistor is '1' the transistor is ON. When the gate of transistor is '0' the transistor is OFF.

The switch level model with transistor symbol is shown in Fig. 4.1.

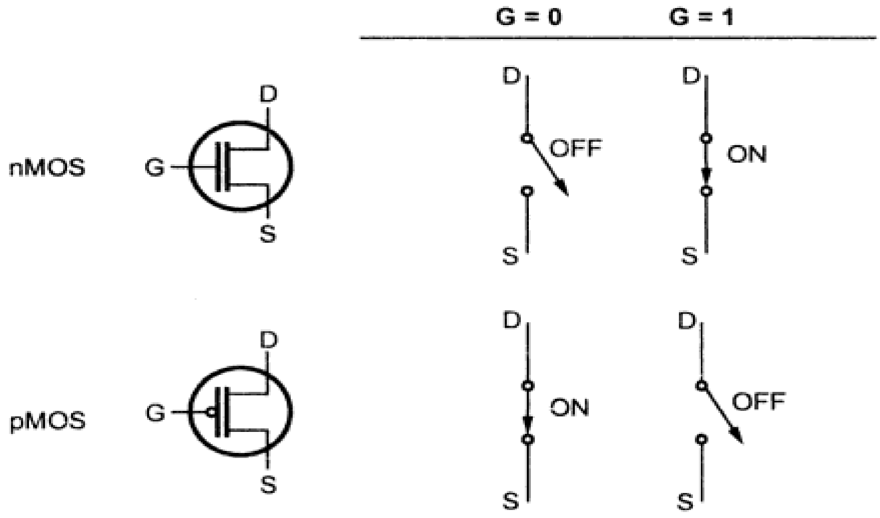


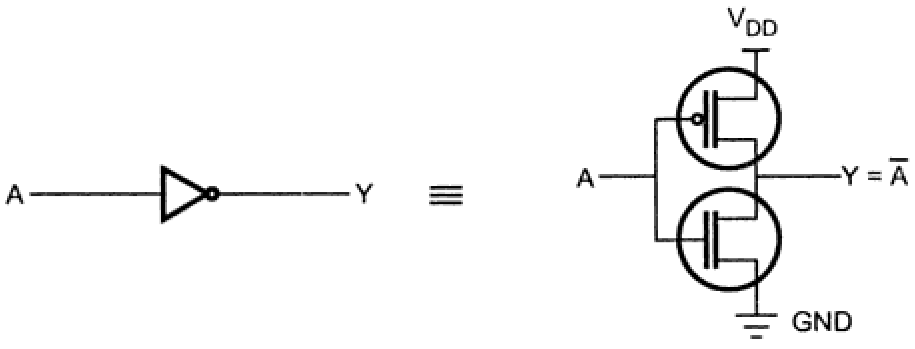
Fig. 4.1 MOSFET as switches

Basic Logic Gates in CMOS

Basic logic gates can be implemented using MOSFETs. Switching property of MOSFET makes it most convenient to implement basic logic gates.

Inverter

A CMOS inverter or NOT gate can be implemented by using one nMOS transistor and one pMOS transistor as shown in Fig. 4.2.



(a) Symbol

(b) Schematic

Fig. 4.2 CMOS inverter

When the input A is '0' , the nMOS transistor is OFF and pMOS transistor is ON. Thus the output Y is pulled to '1' as it is connected to V_{DD}. Similarly when input A is '1', the pMOS transistor is OFF and Y is pulled down to '0'. Truth table is summarized in Table 4.1.

A	Y
0	1
1	0

Table 4.1 Inverter truth table

OR operation

Logical OR operation can be implemented by connecting pMOS and nMOS as shown in Fig. 4.3.

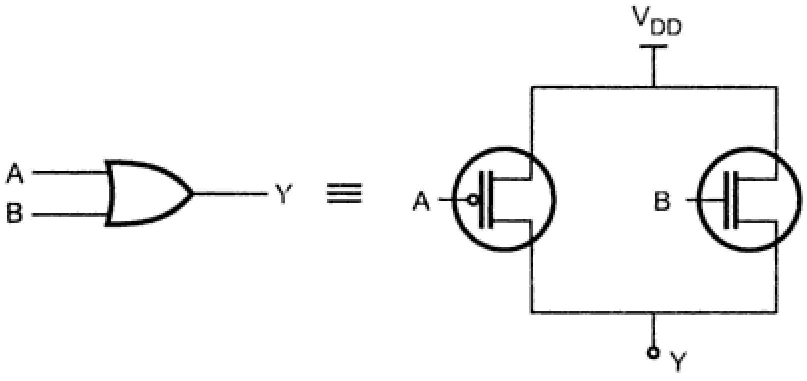


Fig. 4.3 OR operation using CMOS

AND operation

The arrangement for pMOS and nMOS for AND operation is shown in Fig. 4.4.

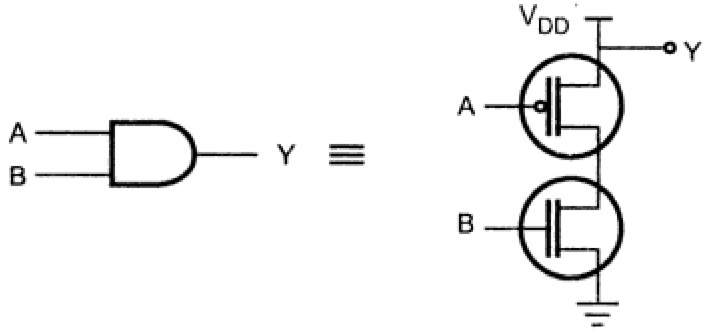


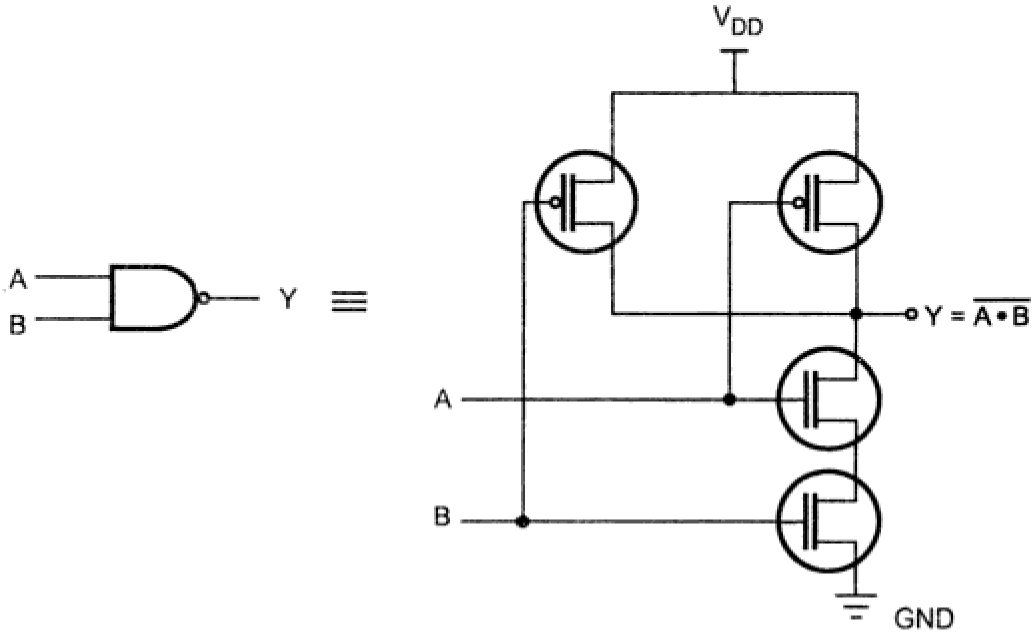
Fig. 4.4 CMOS AND operation

Complex Logic Gate

CMOS can be used to implement other complex logic operations such as two input NAND gate, three input NAND gate, two input NOR, three input NOR and other compound logic operation.

Two Input NAND

Two input NAND consists of two series nMOS transistors between Y and GND and two parallel pMOS transistors between Y and V_{DD} as shown in Fig. 4.5.



(a) Symbol

(b) Schematic

Fig. 4.5 Two input NAND gate

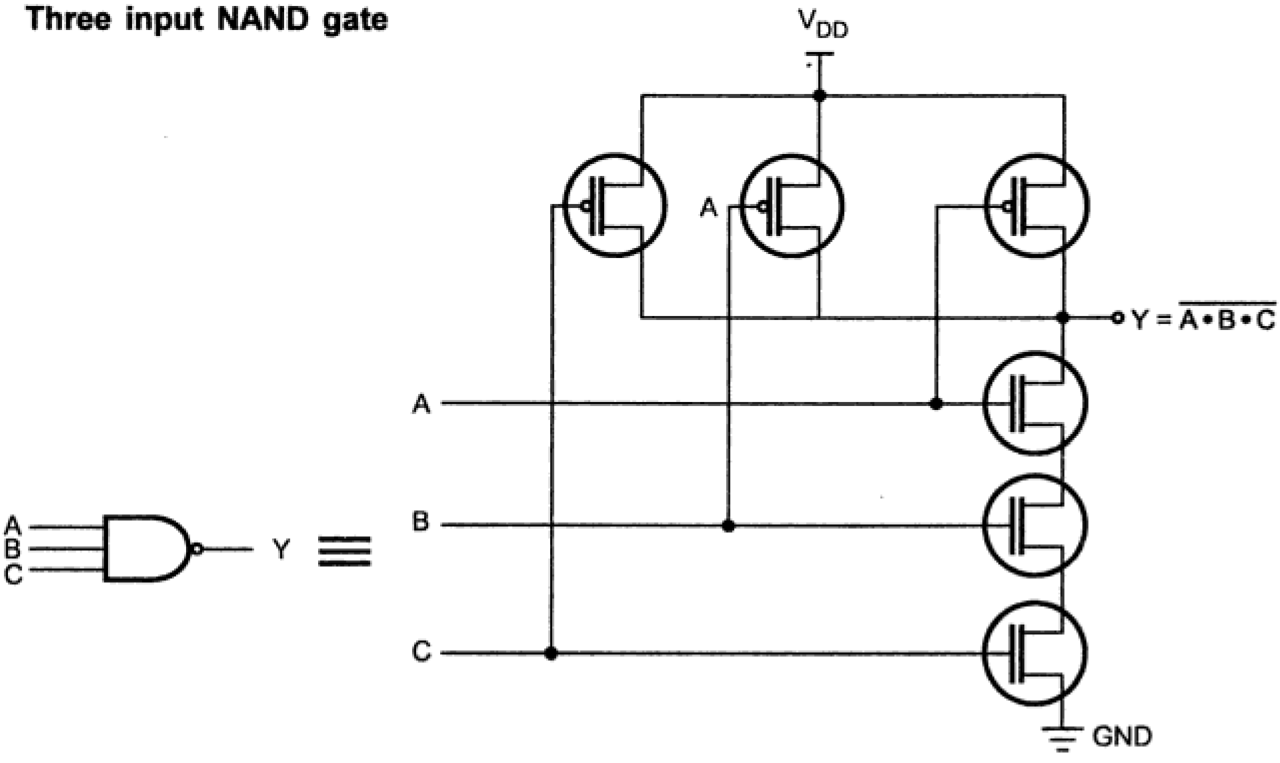
If either input A or B is '0', one of the nMOS transistor will be OFF, the path between Y and GND. At the same time atleast one of pMOS transistor is ON, making a path from Y to V_{DD}. Therefore, the output will be '1'.

If both the inputs are '1', both nMOS transistors will be ON and both pMOS will be OFF, hence the output will be '0'. Truth table is shown in the Table. 4.2.

A	B	Pull-down Network	Pull-up Network	Output (Y)
0	0	OFF	ON	1
0	1	OFF	ON	1
1	0	OFF	ON	1
1	1	ON	OFF	0

Table 4.2 Two input NAND gate truth table

Three input NAND gate



(a) Symbol **(b) Schematic**

Fig. 4.6 Three input NAND gate

When any of the inputs are '0', the output is pulled high through the parallel pMOS transistors. When all of the inputs are '1', the output is pulled low through the series of nMOS transistors.

Combinational Logic

Inverter and NAND gates are examples of complementary CMOS logic gates or static CMOS gates. It has a pull-down network (nMOS) to connect output to '0'(GND) and a pull-up network (pMOS) to connect output to '1'(V_{DD}). The arrangement of network is such that one is ON and other is OFF for any input pattern. Fig. 4.7 shows typical arrangement of pull-up and pull-down networks.

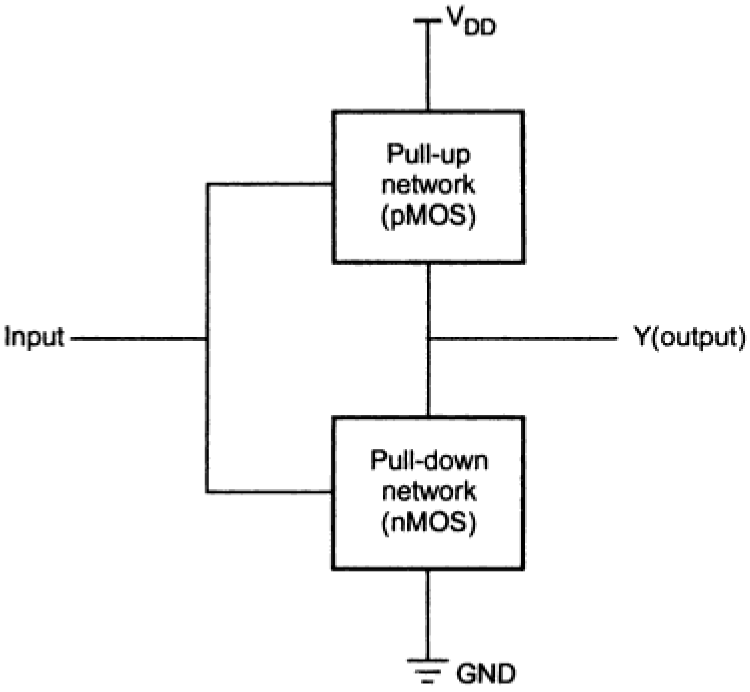


Fig. 4.7 Arrangement of pull-up and pull-down network

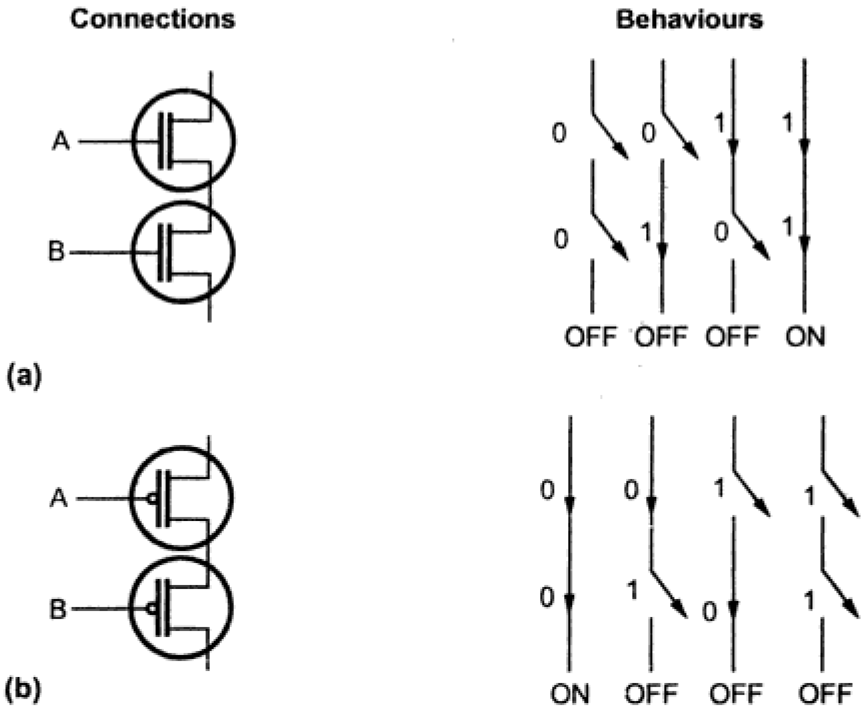
The pull-up and pull-down networks may consists of series or parallel combination of pMOS and nMOS transistors. Two or more transistors in series are ON if all the transistors are ON and the parallel transistors are ON if any of the parallel transistors are ON. By using this principle various combinational logic can be constructed. Selected connections and behaviour of series and parallel transistors are shown in Fig. 4.8.

The output of CMOS logic gate can have four states. Corresponding to various combinations of pull-up and pull-down networks.

	Pull-up OFF	Pull-up ON
Pull-down OFF	Z	1
Pull-down ON	0	Crowbarred (X)

Table 4.3 Output states of CMOS logic

The '1' and '0' levels are the logic states of NAND gate and inverter as either pull-up or pull-down structure is ON. But when both pull-up and pull-down are OFF, the *high impedance* or *floating* Z output state results. The high impedance or floating state is important in multiplexers, memory elements and bus drivers.



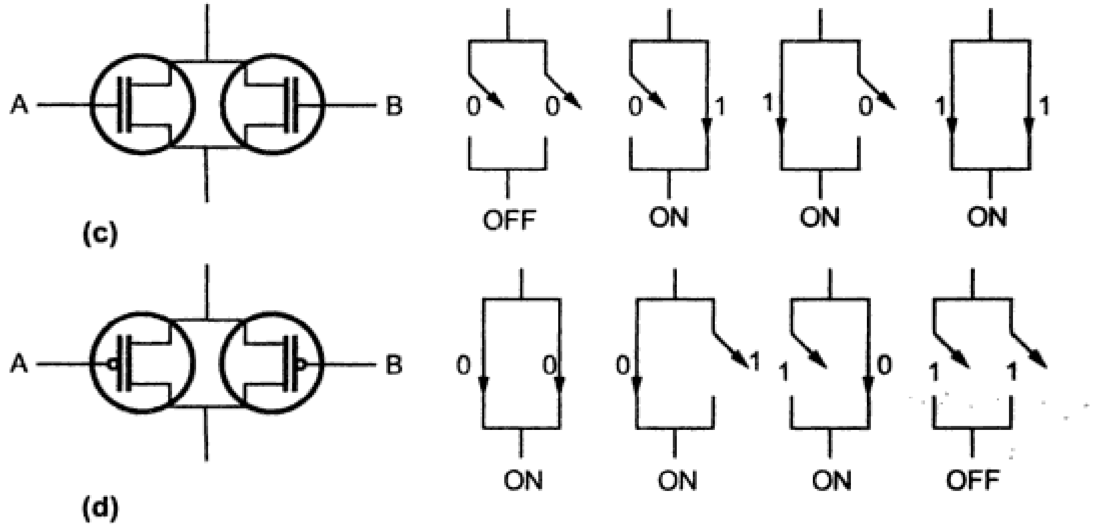


Fig. 4.8 Series and parallel connections of transistors

When both either pull-up or pull-down structures are simultaneously ON a state called *Crowbarred* X level, i.e. intermediate level. This is unwanted situation in CMOS digital circuit.

NOR Gate

Two input CMOS NOR gate is shown in Fig. 4.9.

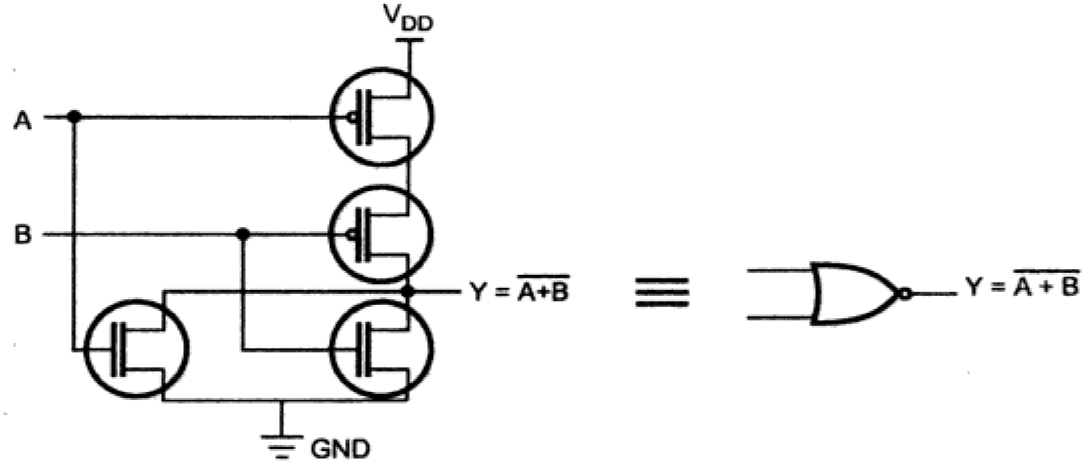


Fig. 4.9 NOR gate

The nMOS transistors are in parallel to pull the output low when any of the input is high. The pMOS transistors are in series to pull the output high when both inputs are low as shown in truth table.

NOR gate truth table		
A	**B**	**Y**
0	0	1
0	1	0
1	0	0
1	1	0

The crowbarred or floating condition does not exist in NOR gate.

Three Input NAND Gate

Fig. 4.10 shows schematic for three input NOR gate.

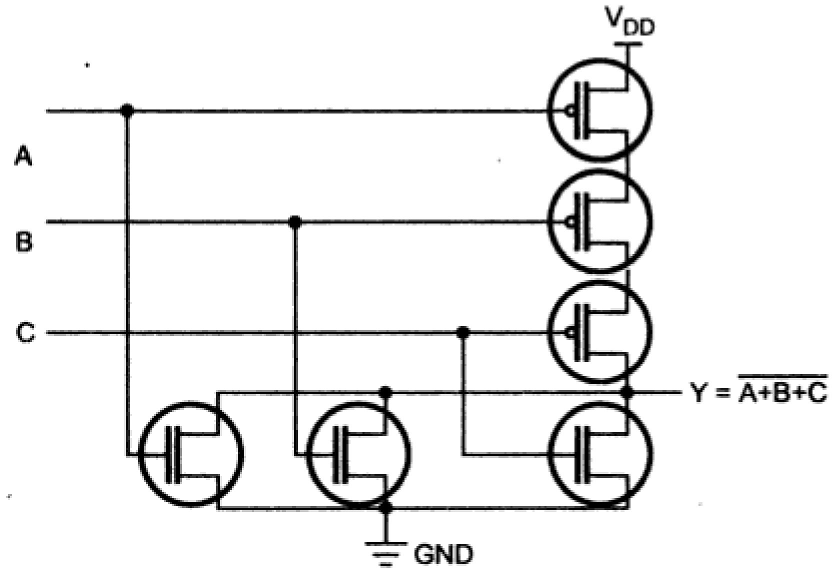
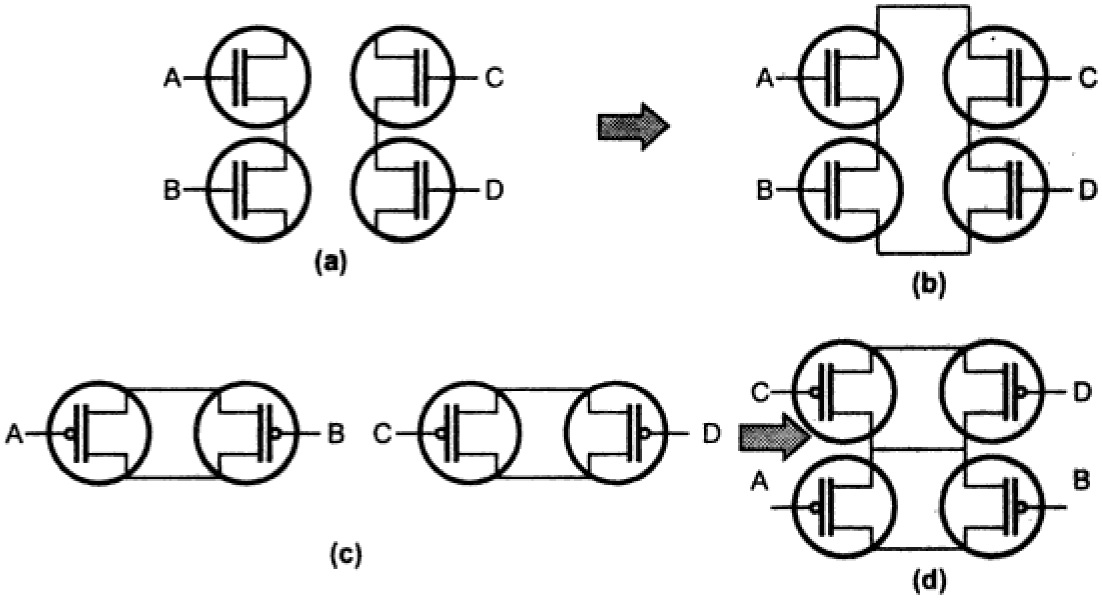


Fig. 4.10 Three input NOR gate

When any of the input is high output is pulled low through parallel nMOS transistors. If all inputs are low, the output is pulled high through the series pMOS transistors.

Compound Gates

The compound gate comprises of combination of series and parallel switch structures. A function $Y=\overline{(A\cdot B)+(C\cdot D)}$ is a good example of compound gate. The function is called AND-OR-INVERT-22 (AOI 22) because it performs the NOR of a pair of two input ANDs. Fig. 4.11 shows implementation of function $Y=\overline{(A\cdot B)+(C\cdot D)}$.

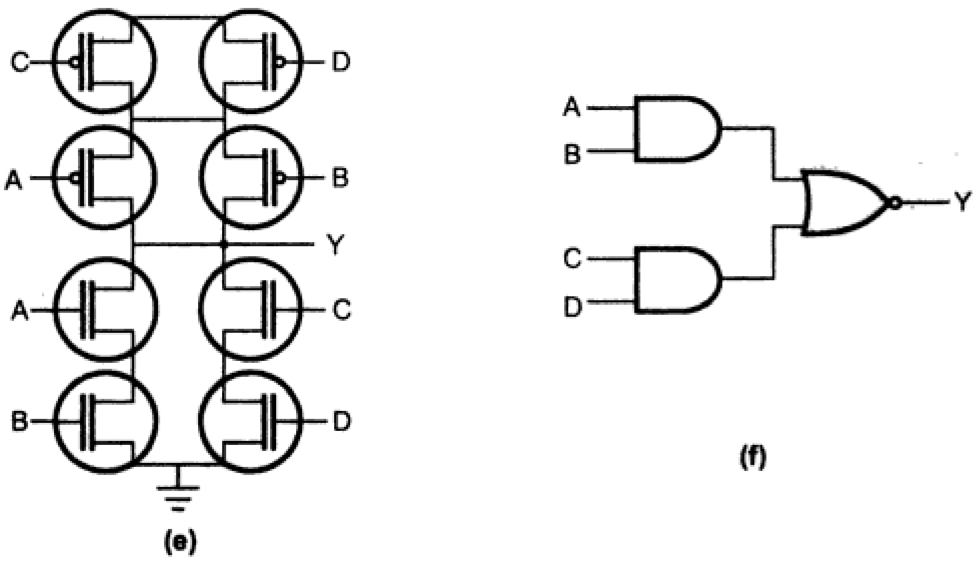


(a)

(b)

(c)

(d)



(e)

(f)

Fig. 4.11 CMOS compound gate of function $Y=\overline{(A \cdot B)+(C \cdot D)}$

Transmission Gates

The strength of signal is measured by how closely it approximates ideal voltage source. A strong signal can sink or source more current. The strongest source of '1's and '0's are power supplies or rails (V_{DD} and GND).

An nMOS transistor passes a strong '0' since it is almost perfect switch when passing a '0'. But it is imperfect at passing a '1'. A pMOS transistor has opposite behaviour i.e. passing a strong '1's and degraded '0's. Fig. 4.12 shows transistor symbols and their behaviour.

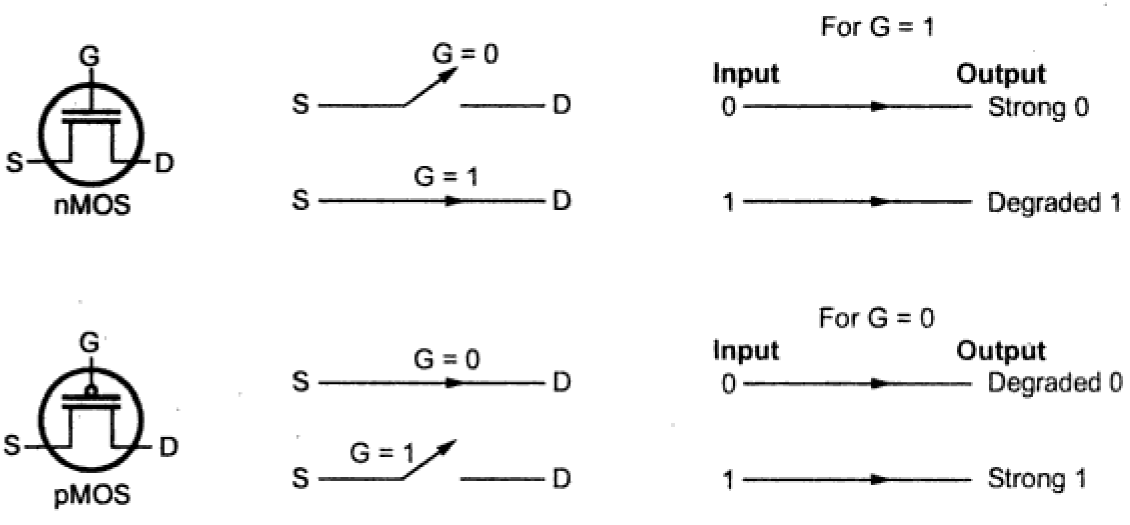


Fig. 4.12 Pass transistor strong and degraded outputs

A pMOS or nMOS transistor alone is (an imperfect switch) called as **pass transistor.** A parallel combination of nMOS and pMOS transistor is called as **transmission gate** or **pass gate**. The transmission gate turns on when a '1' is applied to G in which '0's and '1's are both passed in an acceptable form.

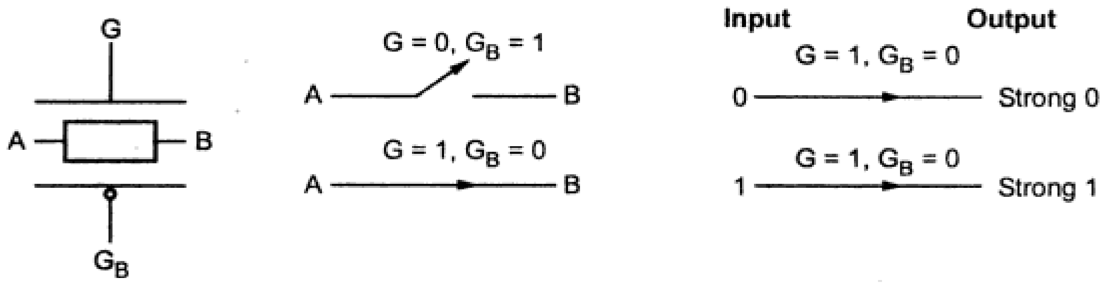


Fig. 4.13 Transmission gate

Multiplexers

Multiplexers are important components in CMOS memory elements and data manipulation structures. A multiplexer selects one of the inputs to the output line depending on select lines.

Let in a two input (2 : 1) MUX the inputs are D_0 and D_1 and select lines are S and $\bar{S}$. The input D_0 is selected when select S = 1 and $\bar{S}$ = 0 and input D_1 is selected when select S = 0 and $\bar{S}$ = 1.

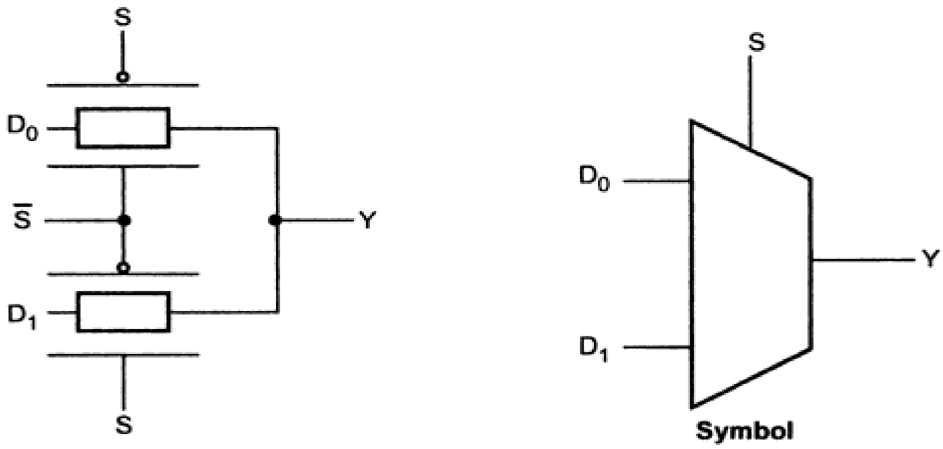
The logic expression is $Y = \bar{S} \cdot D_0 + S \cdot D_1$.

Truth table for two input MUX is shown in Table 4.4.

Inputs		Select lines		Output
D_1	D_0	S	$\bar{S}$	Y
X	0	0	1	0
X	1	0	1	1
0	X	1	0	0
1	X	1	0	1

Table 4.4 Two input MUX truth table

Two input multiplexer can be formed by connecting two transmission gates together as shown in Fig. 4.14. Any one of transmission gate is enabled depending on select line inputs.



Symbol

Fig. 4.14 Transmission gate multiplexer and symbol

The transmission gates make multiplexer non-restoring. A restoring, inverting multiplexer can be built. One method is by using compound gate as shown in Fig. 4.15.

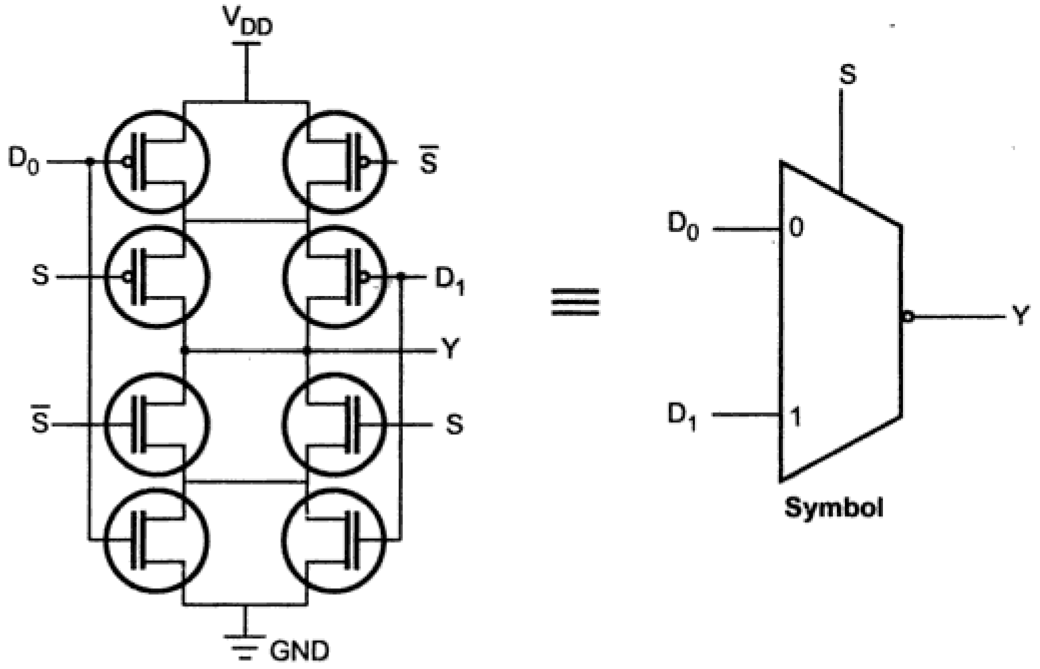


Fig. 4.15 Inverting MUX using compound gate

The other method is to gang together two tristate inverters as shown in Fig. 4.16

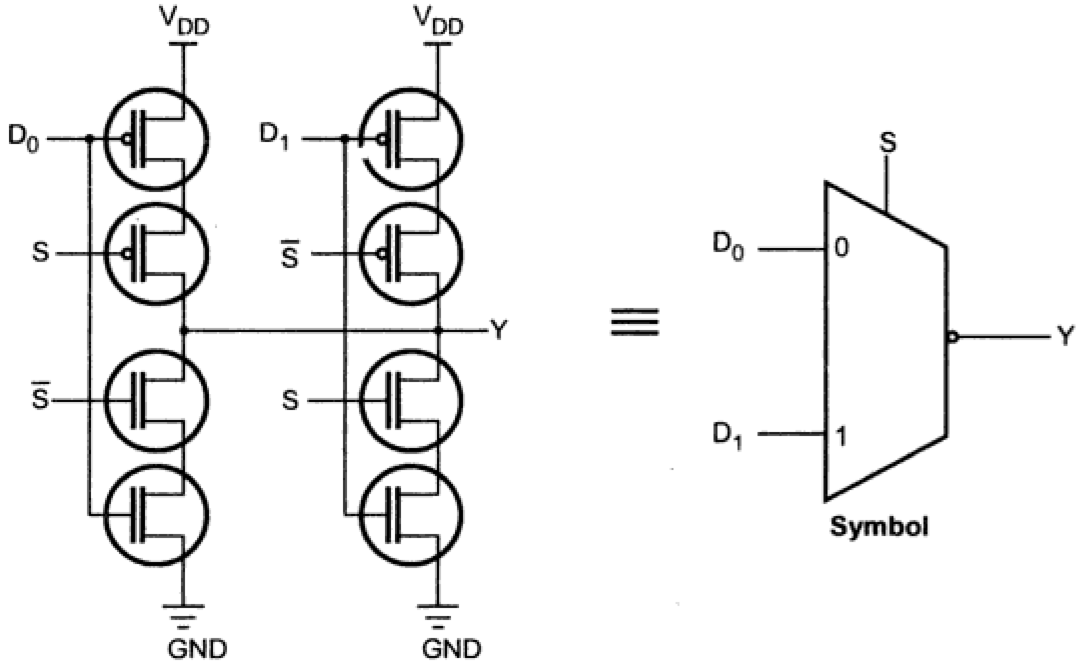


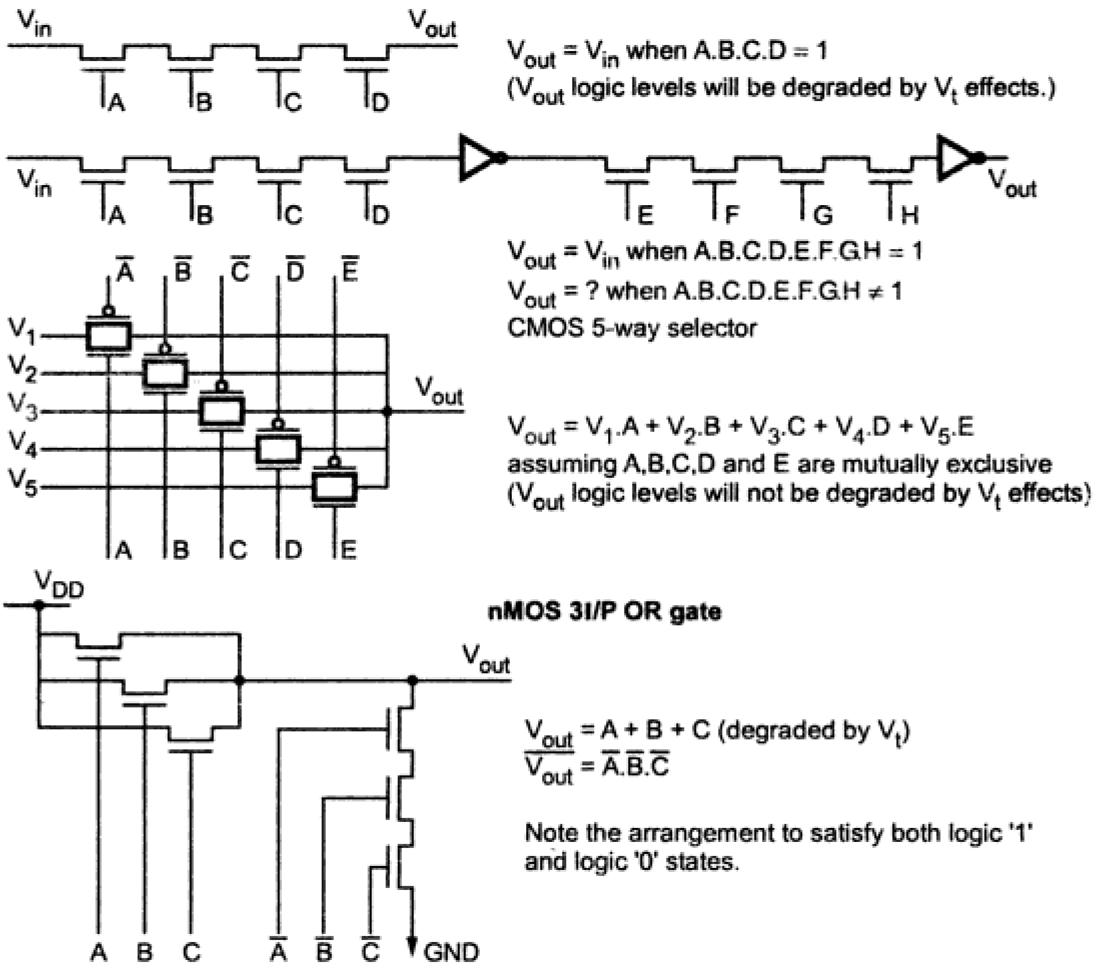
Fig. 4.16 Inverting MUX using tristate inverters

The tristate inverter method is more compact and faster as it requires less internal wire. In the equivalent symbol of MUX the complementary select is generated within cell, hence it is omitted from the symbol.

Switch Logic

The switch logic is built on 'pass transistors' or on transmission gates. It is fast for small arrays and draws no static current from the supply rails. Hence power dissipation of such circuits is small because current flows only on switching.

The pass transistor is used as a switch in relaying the signals. The path through each switch is isolated from the signal activating the switch. This allows the designer to have a considerable freedom in implementing architectural features as compared with bipolar logic-based designs. Switch logic arrangements using basic OR and AND connections are shown in Fig. 4.22 many such combinations of switches are possible.



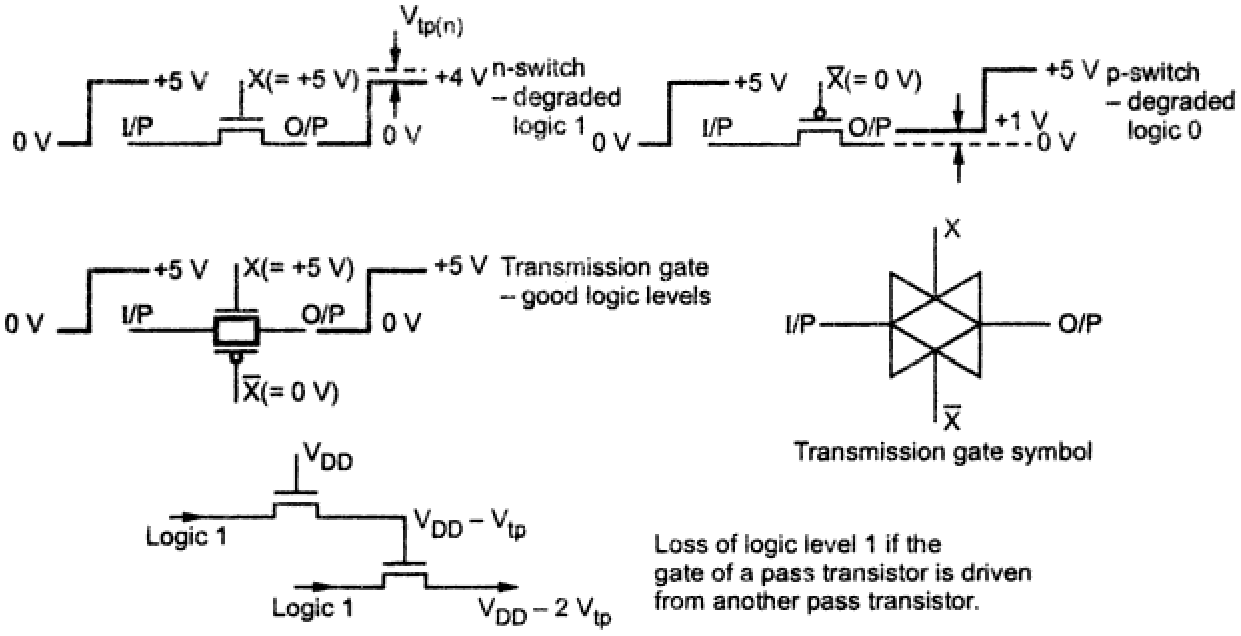
V_{in}
V_{out}
A
B
C
D
$V_{out} = V_{in}$ when A.B.C.D = 1
(V_{out} logic levels will be degraded by V_t effects.)
E
F
G
H
$V_{out} = V_{in}$ when A.B.C.D.E.F.G.H = 1
V_{out} = ? when A.B.C.D.E.F.G.H ≠ 1
CMOS 5-way selector
$\overline{A}$ $\overline{B}$ $\overline{C}$ $\overline{D}$ $\overline{E}$
V_1
V_2
V_3
V_4
V_5
$V_{out} = V_1.A + V_2.B + V_3.C + V_4.D + V_5.E$
assuming A,B,C,D and E are mutually exclusive
(V_{out} logic levels will not be degraded by V_t effects)
V_{DD}
nMOS 3I/P OR gate
$V_{out} = A + B + C$ (degraded by V_t)
$\overline{V_{out}} = \overline{A}.\overline{B}.\overline{C}$
Note the arrangement to satisfy both logic '1' and logic '0' states.
A B C
$\overline{A}$ $\overline{B}$ $\overline{C}$
GND

Fig. 4.22 Some switch logic arrangements

Pass Transistors and Transmission Gates

Switches or switch logic can be realized either using simple n or p pass transistors or from transmission gates. The transmission gates are complementary switches made up of a p-pass and a n-pass transistor in parallel as shown in Fig. 4.23. Simple pass transistors may suffer from undesirable threshold voltage effects which gives rise to loss of logic levels as indicated in Fig. 4.23.

For this reason, the apparently complex transmission gate is preferred to the simple n-switch or p-switch in most CMOS applications. The transmission gate is free from any such degradation of logic levels although it occupies more area and requires complementary signals to drive it. Also, the 'on' resistance of transmission gates is lower than that of the simple pass transistor switches.



Loss of logic level 1 if the gate of a pass transistor is driven from another pass transistor.

Fig. 4.23 Some properties of pass transistors and transmission gates

There is one restriction which must be observed when using nMOS switch logic in the way that no pass transistor gate input must be driven through one or more pass transistors. This restriction is illustrated in Fig. 4.23. As shown here, the logic levels, logic levels propagated through pass transistors get degraded by threshold voltage effects. The signal out of the pass transistor T_1 is not a full logic 1 but rather a voltage that is one transistor threshold below the true logic 1. Hence this degraded voltage does not permit the output of T_2 to reach an acceptable logic 1 level.

Alternate Gate Circuits

The availability of both n- and p-transistors makes it possible for the CMOS designer to explore and exploit various alternatives to inverter-based CMOS logic.

Pseudo-nMOS logic

Clearly, if we replace the depletion mode pull-up transistor of the standard nMOS circuits with a p-transistor with gate connected to V_{SS}, we have a structure similar to the nMOS equivalent. This approach to logic design is illustrated by the three-input *Nand* gate in Figure 9. The circuit arrangements look and behave much like nMOS circuits and appropriate ratio rules must be applied.

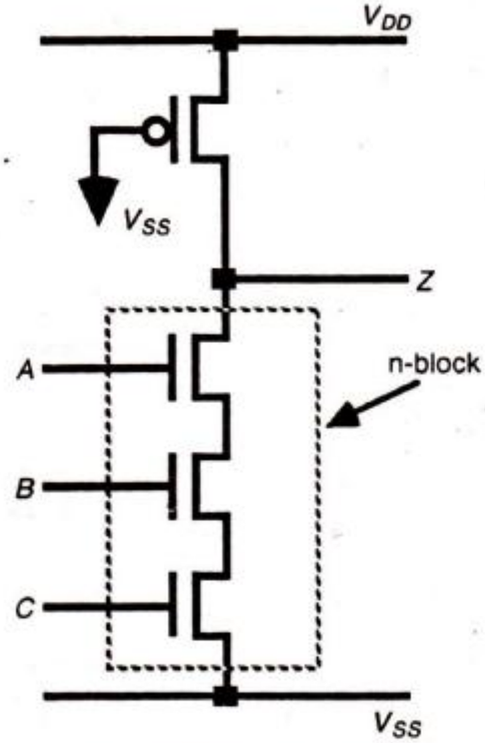


FIGURE .9 Pseudo-nMOS *Nand* gate.

In order to determine the required ratio, we consider the arrangement of Figure 10 in which a pseudo-nMOS inverter is being driven by another similar inverter, and we consider the conditions necessary to produce an output voltage of V_{inv} for an identical input voltage. As for the nMOS analysis, we consider the conditions for which $V_{inv} = V_{DD}/2$.

At this point the n-device is in saturation (i.e. $0 < V_{gsn} - V_{tn} < V_{dsn}$) and the p-device is operating in the resistive region (i.e. $0 < V_{dsp} < V_{gsp} - V_{tp}$). Equating currents of the n-transistor and the p-transistor, and by suitable rearrangement of the resultant expression, we obtain

$$V_{inv} = V_{tn} + \frac{(2\mu_p/\mu_n)^{1/2}[(-V_{DD} - V_{tn})V_{dsp} - V_{dsp}^2]^{1/2}}{(Z_{p.u.}/Z_{p.d.})^{1/2}}$$

where

$$Z_{p.u.} = L_p/W_p$$

and

$$Z_{p.d.} = L_n/W_n$$

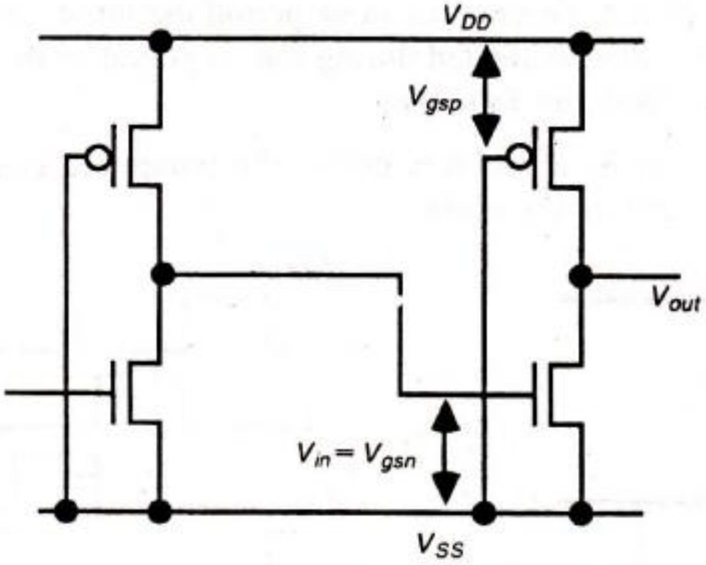


FIGURE 10 Pseudo-nMOS inverter when driven from a similar inverter.

With

$$V_{inv} = 0.5V_{DD}$$

$$V_{tn} = |V_{tp}| = 0.2V_{DD}$$

$$V_{DD} = 5 \text{ V}$$

$$\mu_n = 2.5\ \mu_p$$

we obtain

$$\frac{Z_{p.u.}}{Z_{p.d.}} = \frac{3}{1}$$

A transfer characteristic, V_{out} vs V_{in}, can be drawn and, as for the nMOS case, the characteristic will shift with changes of $Z_{p.u.}/Z_{p.d.}$ ratio.

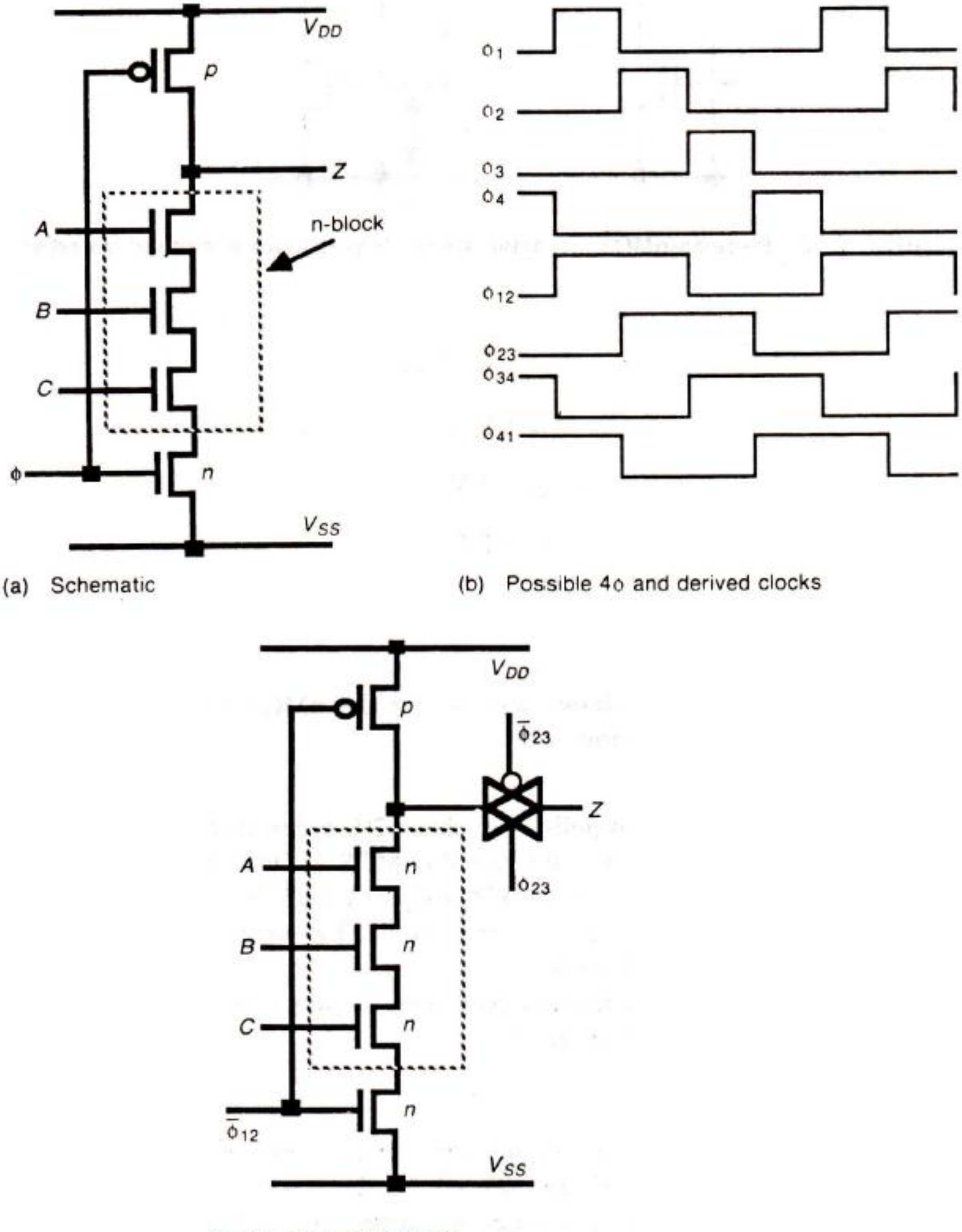
Two points require comment:

1. Since the channel sheet resistance of the p-pull-up is about 2.5 times that of the n-pull-down, and allowing for the ratio of 3:1, the pseudo-nMOS inverter presents a resistance between V_{DD} and V_{SS} which is, say, 85 kΩ compared with 50 kΩ for a comparable 4:1 nMOS device. Thus, power dissipation is reduced to about 60% of that associated with the comparable nMOS device.
2. Owing to the higher pull-up resistance, the inverter pair delay is larger by a factor of 8.5:5 than the 4:1 minimum size nMOS inverter.

Dynamic CMOS logic

The actual logic (see Figure 11(a) for the schematic arrangement) is implemented in the inherently faster nMOS logic (the n-block); a p-transistor is used for the non-time-critical precharging of the output line 'Z' so that the output capacitance is charged to V_{DD} during the off period of the clock signal ϕ. During this same period the inputs are applied to the n-block and the state of the logic is then evaluated during the on period of the clock when the bottom n-transistor is turned on. Note the following:

1. Charge sharing may be a problem unless the inputs are constrained not to change during the on period of the clock.



(a) Schematic

(b) Possible 4ϕ and derived clocks

(c) Type 3 arrangement

FIGURE 11 Dynamic CMOS logic three-input *Nand* gate.

2. *Single phase dynamic logic structures cannot be cascaded* since, owing to circuit delays, an incorrect input to the next stage may be present when evaluation begins, so that its output is inadvertently discharged and the wrong output results.

One remedy is to employ a four-phase clock in which the actual signals used are the derived clocks ϕ_{12}, ϕ_{23}, ϕ_{34}, and ϕ_{41}, as illustrated in Figure 11(b).

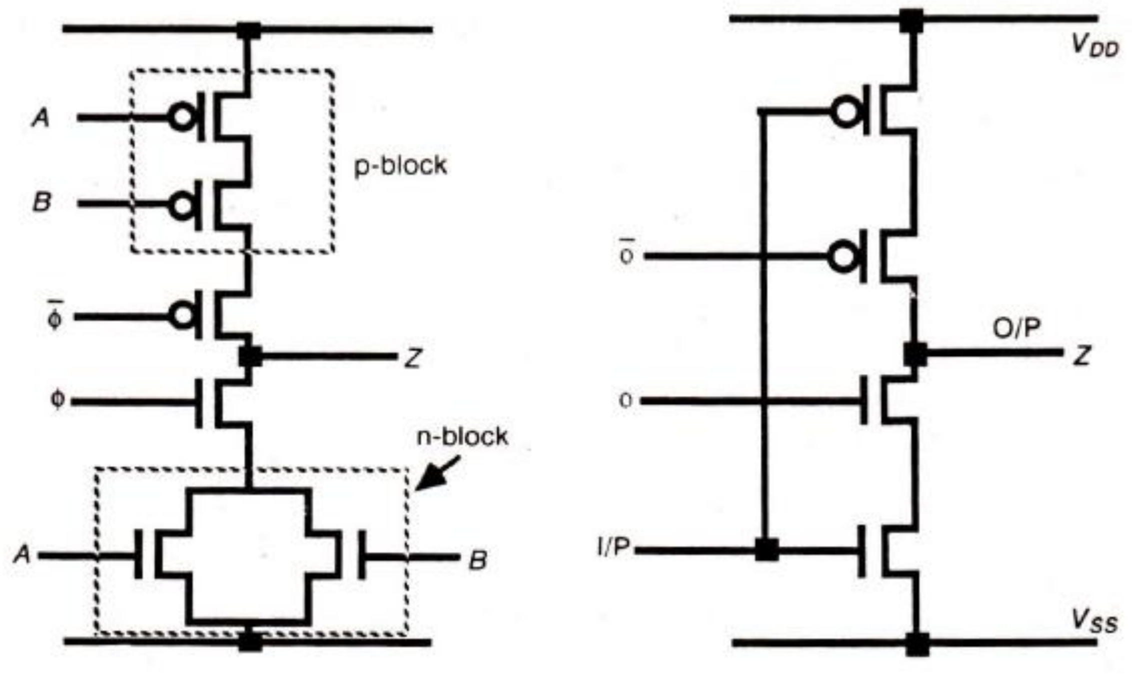
The basic circuit of Figure 11(a) is modified by the inclusion of a transmission gate as in Figure 11(c), the function of which is to sample the output during the 'evaluate' period and to hold the output state while the next stage logic evaluates. For this strategy to work, the next stage must operate on overlapping but later clock signals. Clearly, since there are four different derived clock signals which are used in sequential pairs (e.g. ϕ_{12} and ϕ_{23} in Figure 11(c)), there are four different gate clocking configurations. These configurations are usually identified by a type number which reflects the last of the clock periods activating the gate. For example, the gate shown would be identified as 'type 3' since the output Z is precharged during ϕ_2 and is evaluated during ϕ_3 (the transmission gate is clocked by ϕ_{23}). In order to avoid erroneous evaluations, the gates must be connected in allowable sequences as set out in Table 1.

TABLE 1 Dynamic logic types and sequences

Gate type	*Evaluate clock*	*Transmission gate clock*	*Allowable next types*
Type 1	$\bar{\phi}_{34}$	ϕ_{41}	Types 2 or 3
Type 2	$\bar{\phi}_{41}$	ϕ_{12}	Types 3 or 4
Type 3	$\bar{\phi}_{12}$	ϕ_{23}	Types 4 or 1
Type 4	$\bar{\phi}_{23}$	ϕ_{34}	Types 1 or 2

Clocked CMOS (C^2MOS) logic

The general arrangement may be made clearer by Figure 12. The logic is implemented in both n- and p-transistors in the form of a pull-up p-block and a complementary n-block pull-down structure (Figure 12(a)), as for the inverter-based CMOS logic discussed earlier. However, the logic in this case is evaluated (connected to the output) only during the on period of the clock. As might be expected, a clocked inverter circuit forms part of this family of logic as shown in Figure 12(b). Owing to the extra transistors in series with the output, slower rise-times and fall-times can be expected.



(a) 2 1/P *Nor* gate

(b) Inverter

FIGURE 12 Clocked CMOS (C^2MOS) logic.

CMOS domino logic

An extension to the dynamic CMOS logic discussed earlier is set out in Figure 13. This modified arrangement allows for the cascading of logic structures using only a single phase clock. This requires a static CMOS buffer in each logic gate.

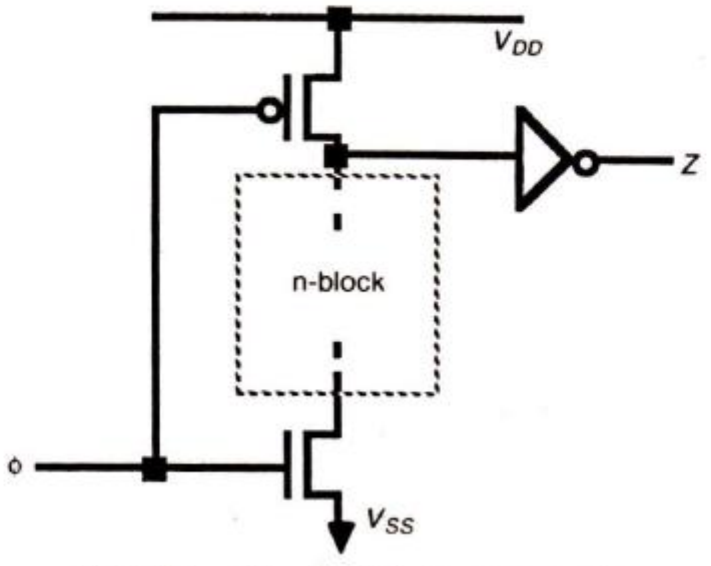


FIGURE 13 CMOS domino logic.

The following remarks will help to place this type of logic in the scheme of things:

1. Such logic structures can have smaller areas than conventional CMOS logic.
2. Parasitic capacitances are smaller so that higher operating speeds are possible.
3. Operation is free of glitches since each gate can make only one '1' to '0' transition.
4. Only non-inverting structures are possible because of the presence of the inverting buffer.
5. Charge distribution may be a problem and must be considered.

n-p CMOS logic

This is another variation of basic dynamic logic arrangement, in which the actual logic blocks are alternately 'n' and 'p' in a cascaded structure as in Figure 14. The precharge and evaluate transistors are fed from the clock ϕ and clockbar $\overline{\phi}$ alternately, and clearly the functions of the top and bottom transistors also alternate, between precharge and evaluate.

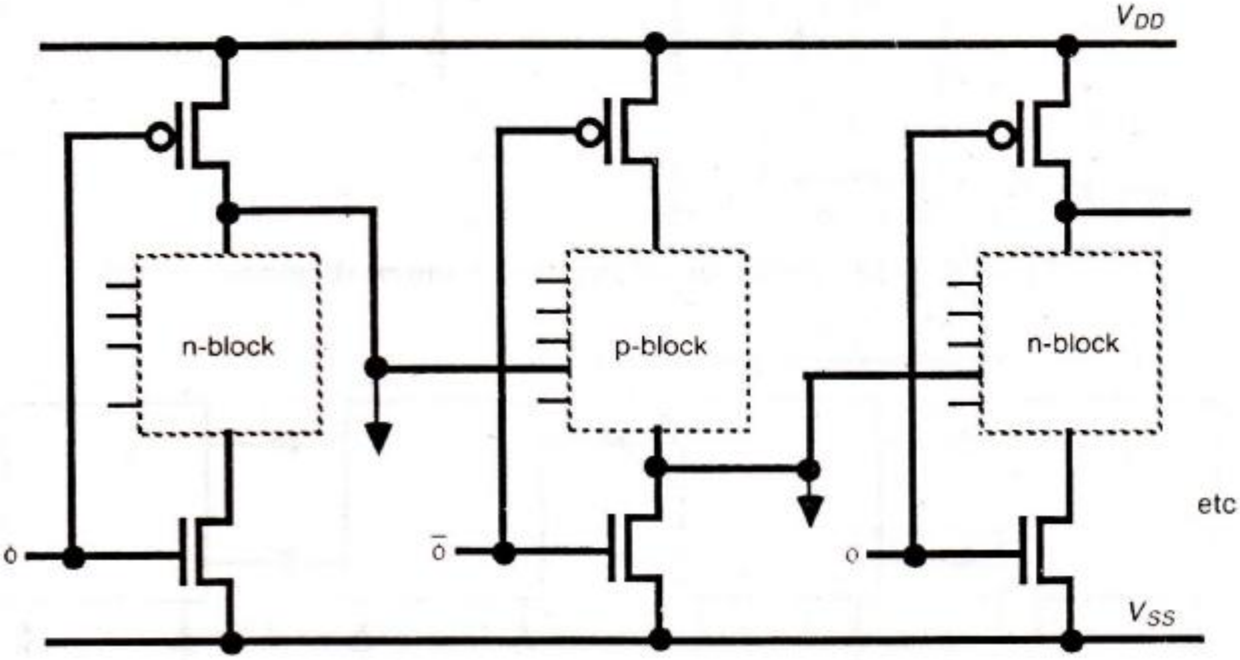


FIGURE 14 n-p CMOS logic.

Basic Circuit Concepts

The active devices of MOS technology having been dealt with in some measure, it is now appropriate to consider their interconnection as circuits. The 'wiring-up' of circuits takes place through the various conductive layers which are produced by the MOS processing and it is therefore necessary to be aware of the resistive and capacitive characteristics of each layer.

Concepts such as sheet resistance R_s and a standard unit of capacitance □*Cg*, help greatly in evaluating the effects of wiring and input and output capacitances. Further, the delays associated with wiring, with inverters and with other circuitry may be conveniently evaluated in terms of a delay unit τ:

SHEET RESISTANCE R_S

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 4.1.

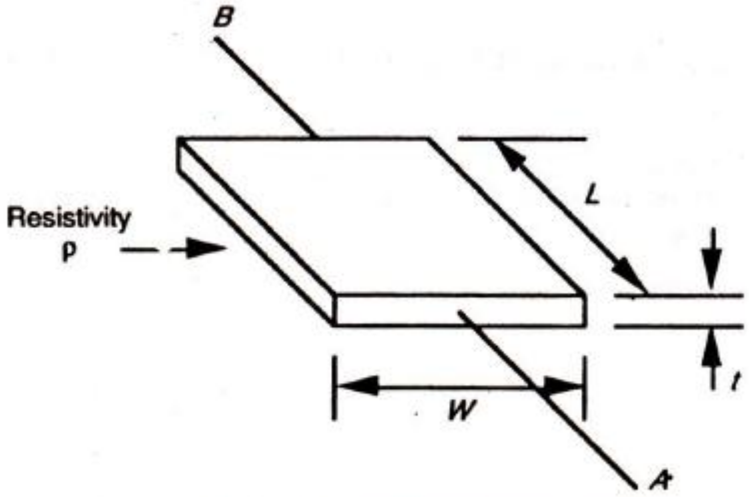


FIGURE 4.1 Sheet resistance model.

With reference to Figure 4.1, 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

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

Note that R_s is completely independent of the area of the square; for example, a 1 μm per side square slab of material has exactly the same resistance as a 1 cm per side square slab of the same material if the thickness is the same.

Thus the actual values associated with the layers in a MOS circuit depend on the thickness of the layer and the resistivity of the material forming the layer. For the metal and polysilicon layers, the thickness of a layer is easily envisaged and the resistivity of the material is known. For the diffusion layer, the depth of the diffusion regions contributes toward the effective thickness while the impurity concentration (or doping level) profile determines the resistivity.

For the MOS processes considered here, typical values of sheet resistance are given in Table 4.1.

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	$2.5 \times 10^{4\dagger}$	$4.5 \times 10^{4\dagger}$	$4.5 \times 10^{4\dagger}$

SHEET RESISTANCE CONCEPT APPLIED TO MOS TRANSISTORS AND INVERTERS

Consider the transistor structures of Figure 4.2 and note that the diagrams distinguish the actual diffusion (active) regions from the channel regions. (*Note:* From here on, the term 'diffusion' also covers active regions in Orbit processes.) The thinox mask layout is the union of diffusion and channel regions and these regions have differing hatching patterns to stress the fact that the polysilicon and underlying silicon dioxide mask the substrate so that diffusion takes place only in the areas defined by the thinox mask which do not coincide with the polysilicon mask.

The simple n-type pass transistor of Figure 4.2(a) has a channel length $L = 2\lambda$ and a channel width $W = 2\lambda$. The channel is, therefore, square and channel resistance (with or without implant)

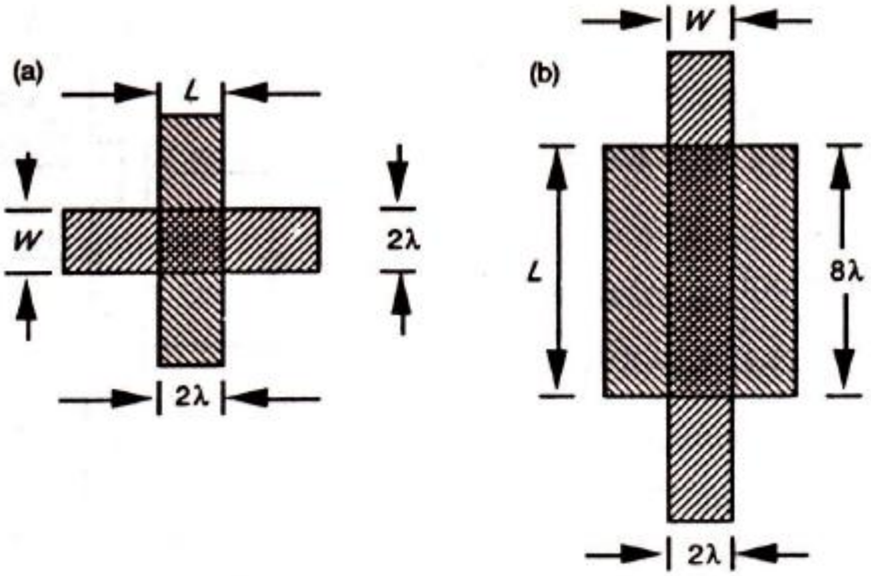


FIGURE 4.2 Resistance calculation for transistor channels.

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

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,

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

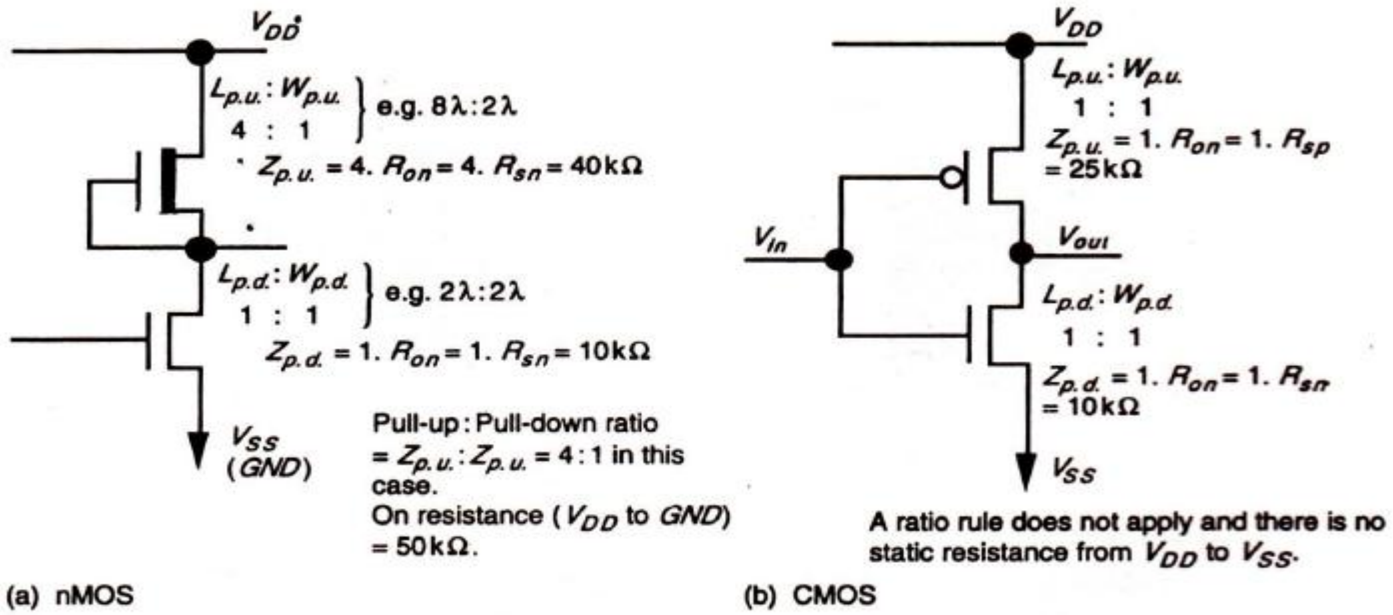
Thus, channel resistance

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

Another way of looking at this is to recognize that this channel can be regarded as four $2\lambda \times 2\lambda$ squares in series, thus giving a resistance of $4R_s$. This particular way of approaching the calculation of resistance is often useful, particularly when dealing with shapes which are not simple rectangles.

Figure 4.3 takes these considerations one step further and shows how the pull-up to pull-down ratio of an inverter is determined. In the nMOS case a simple 4:1 $Z_{p.u.}:Z_{p.d.}$ ratio obviously applies. Note, for example, that a 4:1 ratio would also be achieved if the upper channel (p.u.) length $L = 4\lambda$, and width $W = 2\lambda$ with lower channel (p.d.) length $L = 2\lambda$, and width $W = 4\lambda$.

For the CMOS case, note the different value of R_s which applies for the pull-up transistor.


V_{DD}
$L_{p.u.}:W_{p.u.}$ 4 : 1
e.g. 8λ:2λ
$Z_{p.u.}$ = 4. R_{on} = 4. R_{sn} = 40kΩ
$L_{p.d.}:W_{p.d.}$ 1 : 1
e.g. 2λ:2λ
$Z_{p.d.}$ = 1. R_{on} = 1. R_{sn} = 10kΩ
V_{SS} (GND)
Pull-up: Pull-down ratio = $Z_{p.u.}:Z_{p.u.}$ = 4:1 in this case.
On resistance (V_{DD} to GND) = 50kΩ.
(a) nMOS
V_{DD}
$L_{p.u.}:W_{p.u.}$ 1 : 1
$Z_{p.u.}$ = 1. R_{on} = 1. R_{sp} = 25kΩ
V_{in}
V_{out}
$L_{p.d.}:W_{p.d.}$ 1 : 1
$Z_{p.d.}$ = 1. R_{on} = 1. R_{sn} = 10kΩ
V_{SS}
A ratio rule does not apply and there is no static resistance from V_{DD} to V_{SS}.
(b) CMOS

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

FIGURE 4.3 Inverter resistance calculation.

AREA CAPACITANCES OF LAYERS

From the diagrams we have used to illustrate the structure of transistors, and from discussions of the fabrication processes, it will be apparent that conducting layers are separated from the substrate and each other by insulating (dielectric) layers, and thus parallel plate capacitive effects must be present and must be allowed for.

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$_2$ ∻ 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/μm^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 4.2 Typical area capacitance values for MOS circuits

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)

STANDARD UNIT OF CAPACITANCE $\square C_g$

It is convenient to employ a standard unit of capacitance that can be given a value appropriate to the technology but can also be used in calculations without associating it with an absolute value. The unit is denoted $\square C_g$ 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 4.2(a), for example, for lambda-based rules.

$\square C_g$ 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

THE DELAY UNIT τ

We have developed the concept of sheet resistance R_s and standard gate capacitance unit $\square C_g$. 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 R_s for an nMOS pass transistor channel), as in Figure 4.6, we have:

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

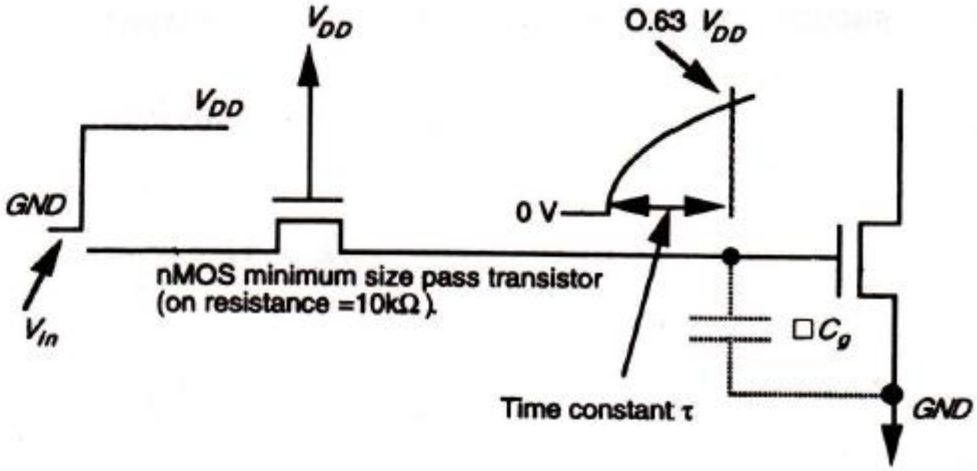


FIGURE 4.6 Model for derivation of τ.

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

$\tau = 0.2$ 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,

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

$$= 0.13 \text{ nsec}$$

This is very close to the theoretical time constant τ calculated above.

Since the transition point of an inverter or gate is 0.5 V_{DD}, which is close to 0.63 V_{DD}, it appears to be common practice to use transit time and time constant (as defined for the delay unit τ) interchangeably and 'stray' capacitances are usually allowed for by doubling (or more) the theoretical values calculated.

In view of this, τ is used as the fundamental time unit and all timings in a system can be assessed in relation to τ.

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 technølogy, $\tau = 0.1$ nsec is also a safe figure.

INVERTER DELAYS

Consider the basic 4:1 ratio nMOS inverter. In order to achieve the 4:1 $Z_{p.u.}$ to $Z_{p.d.}$ ratio, $R_{p.u.}$ will be 4 $R_{p.d.}$ and if $R_{p.d.}$ is contributed by the minimum size transistor then, clearly, the resistance value associated with $R_{p.u.}$ is

$$R_{p.u.} = 4R_s = 40\ \text{k}\Omega$$

Meanwhile, the $R_{p.d.}$ value is $1R_s = 10\ \text{k}\Omega$ so that the delay associated with the inverter will depend on whether it is being turned on or off.

However, if we consider a pair of *cascaded inverters*, then the delay over the pair will be constant irrespective of the sense of the logic level transition of the input to the first. This is clearly seen from Figure 4.7 and, assuming $\tau = 0.3$ nsec and making no extra allowances for wiring capacitance, we have an overall delay of $\tau + 4\tau = 5\tau$. In general terms, the 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.

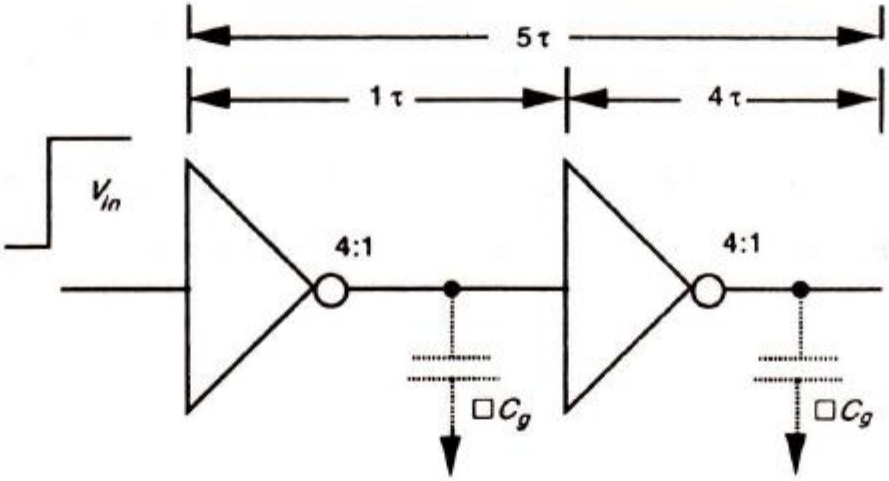


FIGURE 4.7 nMOS inverter pair delay.

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.

When considering CMOS inverters, the nMOS ratio rule no longer applies, but we must allow for the natural (R_s) asymmetry of the usually equal size pull-up p-transistors and the n-type pull-down transistors. Figure 4.8 shows the theoretical delay associated with a pair of minimum size (both n- and p-transistors) lambda-based inverters. Note that the gate capacitance ($= 2\square C_g$) is double that of the comparable nMOS inverter since the input to a CMOS inverter is connected to *both* transistor gates. Note also the allowance made for the differing channel resistances.

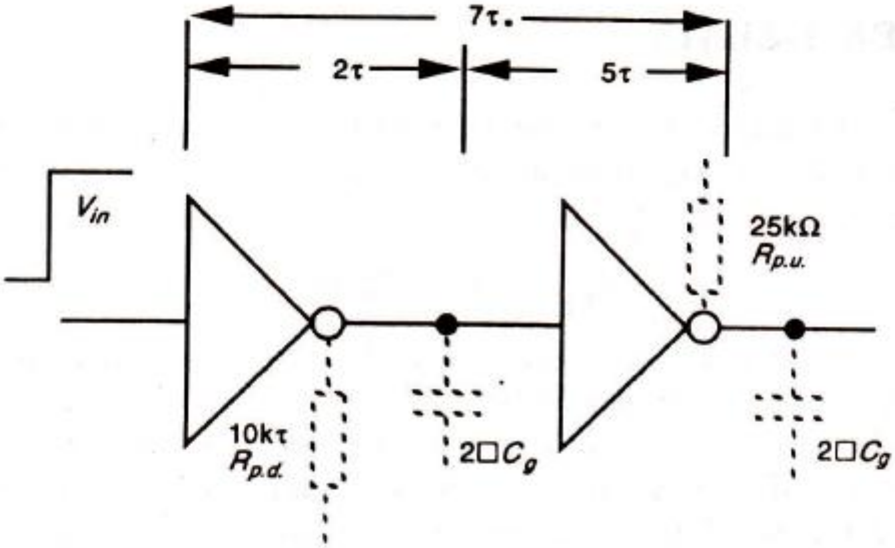


FIGURE 4.8 Minimum size CMOS inverter pair delay.

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

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.

Rise-time estimation

In this analysis we assume that the p-device stays in saturation for the entire charging period of the load capacitor C_L. The circuit may then be modeled as in Figure 4.9.

The saturation current for the p-transistor is given by

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

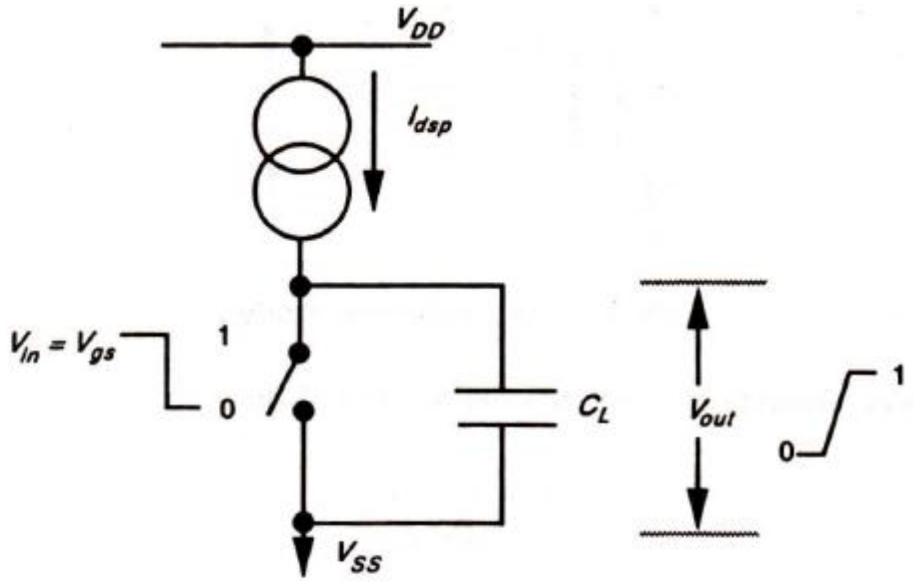


FIGURE 4.9 Rise-time model.

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

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

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 4.10.

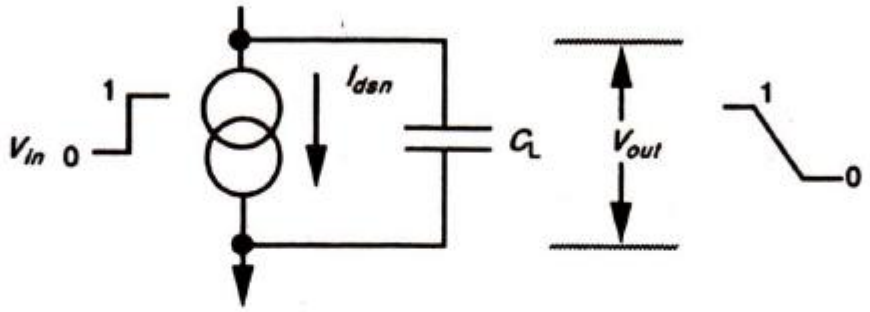


FIGURE 4.10 Fall-time model.

Making similar assumptions we may write for fall-time:

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

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'.

In order to achieve symmetrical operation using minimum channel length, we would need to make $W_p = 2.5W_n$ and for minimum size lambda-based geometries this would result in the inverter having an input capacitance of $1\square C_g$ (n-device) + $2.5\square C_g$(p-device) = $3.5\square C_g$ in total.

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.

DRIVING 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 $\square C_g$ 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

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 $Z_{p.d.}$ and $Z_{p.u.}$ 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 $L \times W$ becomes significant and a comparatively large capacitance is presented at the input, which in turn 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 4.11 (showing nMOS inverters, for example).

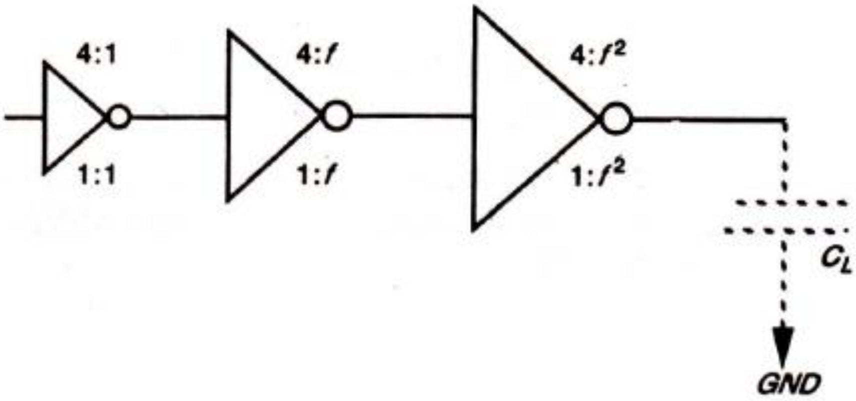


FIGURE 4.11 Driving large capacitive loads.

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 (this treatment is attributed to Mead and Conway).

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

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

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

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

and overall delay t_d

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

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

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

or

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

Super Buffers

The asymmetry of the conventional inverter is clearly undesirable, and gives rise to significant delay problems when an inverter is used to drive more significant capacitive loads.

A common approach used in nMOS technology to alleviate this effect is to make use of super buffers as in Figures 4.12 and 4.13.

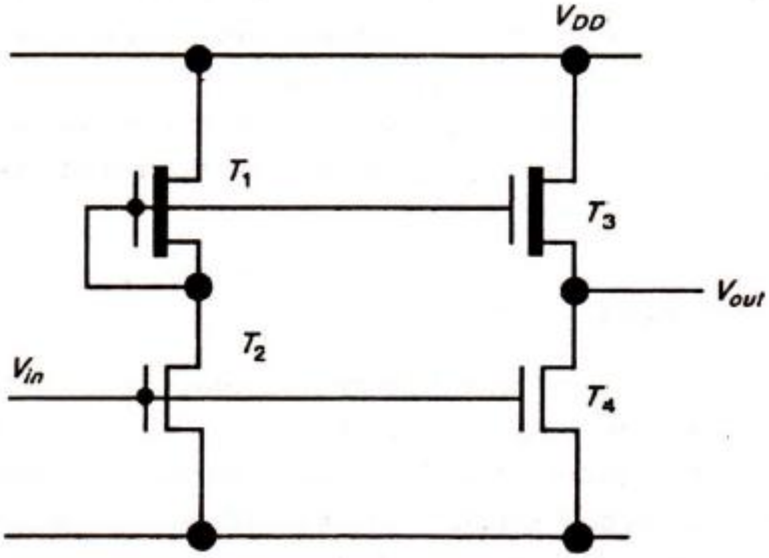


FIGURE 4.12 Inverting type nMOS super buffer.

An inverting type is shown in Figure 4.12; considering a positive going logic transition V_{in} at the input, it will be seen that the inverter formed by T_1 and T_2 is turned on and, thus, the gate of T_3 is pulled down toward 0 volt with a small delay. Thus, T_3 is cut off while T_4 (the gate of which is also connected to V_{in}) is turned on and the output is pulled down quickly.

Now consider the opposite transition: when V_{in} drops to 0 volt, then the gate of T_3 is allowed to rise quickly to V_{DD}. Thus, as T_4 is also turned off by V_{in}, T_3 is made to conduct with V_{DD} on its gate, that is, with twice the average voltage that would apply if the gate was tied to the source as in the conventional nMOS inverter. Now, since I_{ds} α V_{gs} then doubling the effective V_{gs} will increase the current and thus reduce the delay in charging any capacitance on the output, so that more symmetrical transitions are achieved.

The corresponding non-inverting nMOS super buffer circuit is given at Figure 4.13 and, to put matters in perspective, the structures shown when realized in 5 μm technology are capable of driving loads of 2 pF with 5 nsec rise-time.

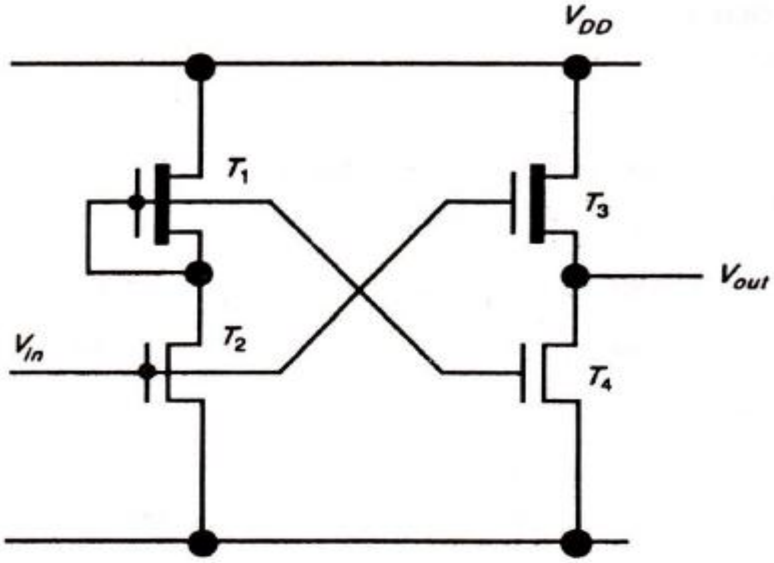


FIGURE 4.13 Non-inverting type nMOS super buffer.

BiCMOS Drivers

The availability of bipolar transistors in BiCMOS technology presents the possibility of using bipolar transistor drivers as the output stage of inverter and logic gate circuits. We have already seen that bipolar transistors have transconductance g_m and current/area I/A characteristics that are greatly superior to those of MOS devices. This indicates high current drive capabilities for small areas in silicon.

Bipolar transistors have an exponential dependence of the output current I_c on the input base to emitter voltage V_{be}. This means that the device can be operated with much smaller input voltage swings than MOS transistors and still switch relatively large currents. Thus, bipolar transistors have a much better switching performance, primarily as a result of the smaller input voltage swings. Only a small amount of charge must be moved during switching.

One point to consider is the possible effect of temperature T on the required input voltage V_{be}. Although V_{be} is logarithmically dependent on base width W_B, doping level N_A, electron mobility μ_n and collector current I_c it is only linearly dependent on T. This means that there is no difficulty in matching V_{be} values across a circuit, spread over an area on chip, as the temperature differences across a chip will not be sufficient to cause more than a few millivolts of difference in V_{be} between any two bipolar transistors.

The switching performance of a transistor driving a capacitive load may be visualized initially from the simple model given in Figure 4.14.

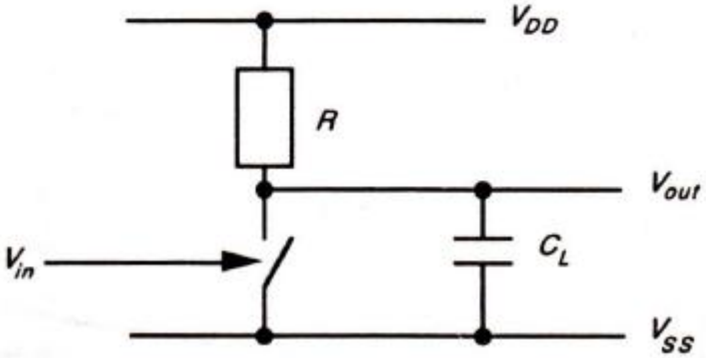


FIGURE 4.14 Driving ability of bipolar transistor.

It may be shown that the time Δt necessary to change the output voltage V_{out} by an amount equal to the input voltage V_{in} is given by

$$\Delta t = \frac{C_L}{g_m}$$

where g_m is the transconductance of the bipolar transistor.

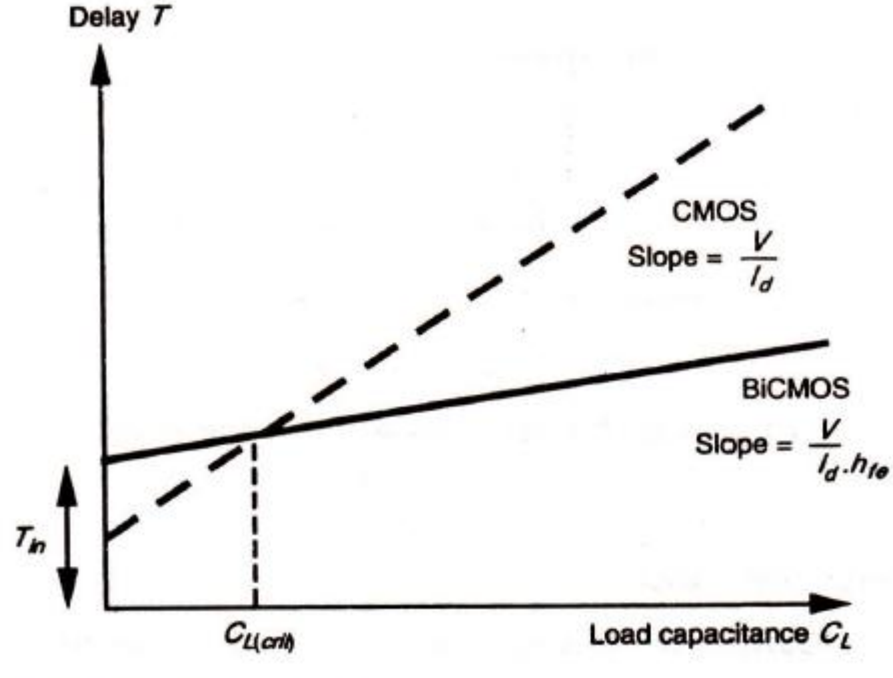
Clearly, since the bipolar transistor has a relatively high transconductance, the value of Δt is small.

A more exacting appraisal of the bipolar transistor delay reveals that it comprises two main components:

1. T_{in}—an initial time necessary to charge the base emitter junction of the bipolar (npn) transistor. Typically, for the BiCMOS transistor-based driver we are considering, T_{in} is in the region of 2ns. A similar consideration of a CMOS transistor driver in the same BiCMOS technology would reveal a figure of 1ns for T_{in}, this being the time taken to charge the input gate capacitance. As a matter of interest, a comparable figure for a GaAs driver is around 50–100 ps.
2. T_L—the time taken to charge the output load capacitance C_L and it will be noted that this time is less for the bipolar driver by a factor of h_{fe}, where h_{fe} is the bipolar transistor gain.

Although the bipolar transistor has a higher value of T_{in}, T_L is smaller because of the faster charging rate as discussed.

The combined effect of T_{in} and T_L is represented in Figure 4.15 and it will be seen that there is a critical value of load capacitance $C_{L(crit)}$ below which the BiCMOS driver is slower than a comparable CMOS driver.



- Delay of BiCMOS inverter can be described by

$$T = T_{in} + (V/I_d)\ (1/h_{fe})\ C_L$$

where

T_{in} = time to charge up base/emitter junction
h_{fe} = transistor current gain (common emitter)

- Delay for BiCMOS inverter is reduced by a factor of h_{fe} compared with a CMOS inverter.

FIGURE 4.15 Delay estimation.

A further significant parameter contributing to delay is the collector resistance R_c of a bipolar transistor. Clearly a high value for R_c will mean a long propagation delay through the transistor when charging a capacitive load. The effect can be assessed from Figure 4.16, which shows typical delay values at two values of C_L for a range of collector resistance R_c. The reason for including the buried subcollector region in the BiCMOS process is to keep R_c as low as possible.

BiCMOS fabrication processes produce reasonably good bipolar transistors—high g_m, high β, high h_{fe} and low R_c—without compromising or overelaborating the basic CMOS process. The availability of bipolar transistors in logic gate and driver/buffer design provides a great deal of scope and freedom for the VLSI designer.

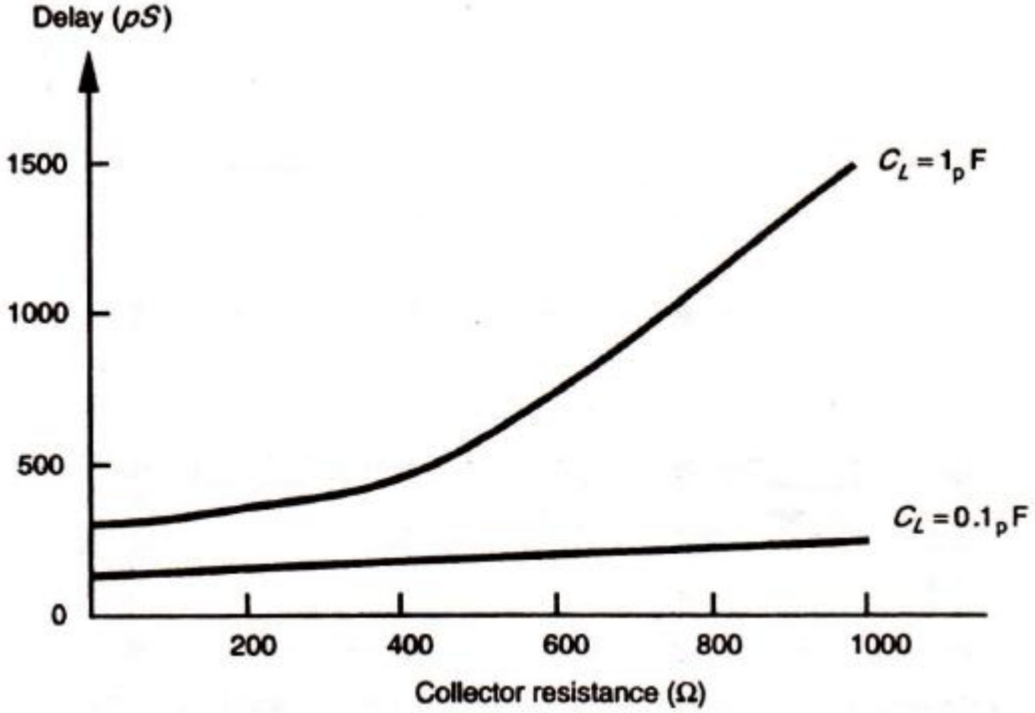


FIGURE 4.16 Gate delay as a function of collector resistance.

PROPAGATION DELAYS

A degree of freedom offered by MOS technology is the use of pass transistors as series or parallel switches in logic arrays. Quite frequently, therefore, logic signals must pass through a number of pass transistors in series. A chain of four such transistors is shown in Figure 4.17(a) in which all gates have been shown connected to V_{DD} (logic 1), which would be the case for a signal to be propagated to the output. The circuit thus formed may be modeled as in Figure 4.17(b) and it is then possible to evaluate the delay through the network.

The response at node V_2 with respect to time is given by

$$C\frac{dV_2}{dt} = (I_1 - I_2) = \frac{[(V_1 - V_2) - (V_2 - V_3)]}{R}$$

In the limit as the number of sections in such a network becomes large, this expression reduces to

$$RC\frac{dV}{dt} = \frac{d^2V}{dx^2}$$

where

R = resistance per unit length
C = capacitance per unit length
x = distance along network from input.

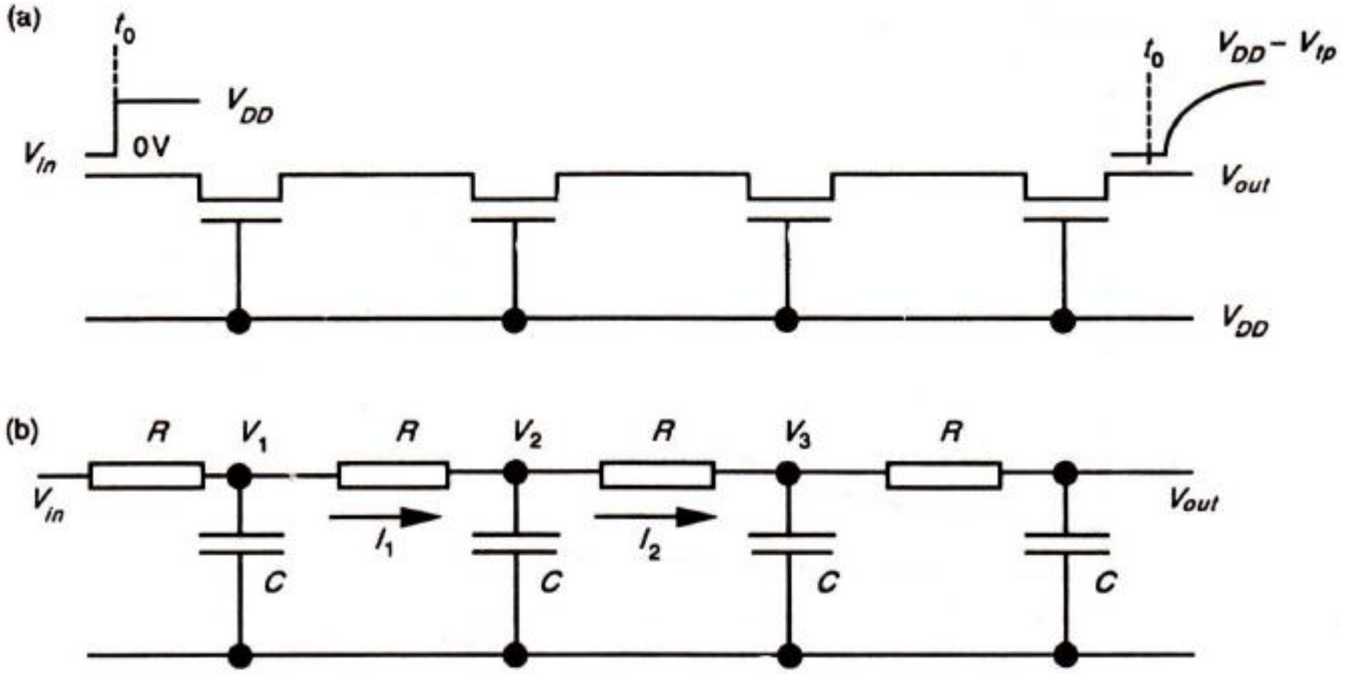


FIGURE 4.17 Propagation delays in pass transistor chain.

The propagation time τ_p for a signal to propagate a distance x is such that

$$\tau_p \alpha x^2$$

The analysis can be simplified if all *Rs* and *Cs* are lumped together, then

$$R_{total} = nrR_s$$

$$C_{total} = nc\square C_g$$

where r gives the relative resistance per section in terms of R_s and c gives the relative capacitance per section in terms of $\square C_g$.

Then, it may be shown that overall delay τ_d for n sections is given by

$$\tau_d = n^2 rc(\tau)$$

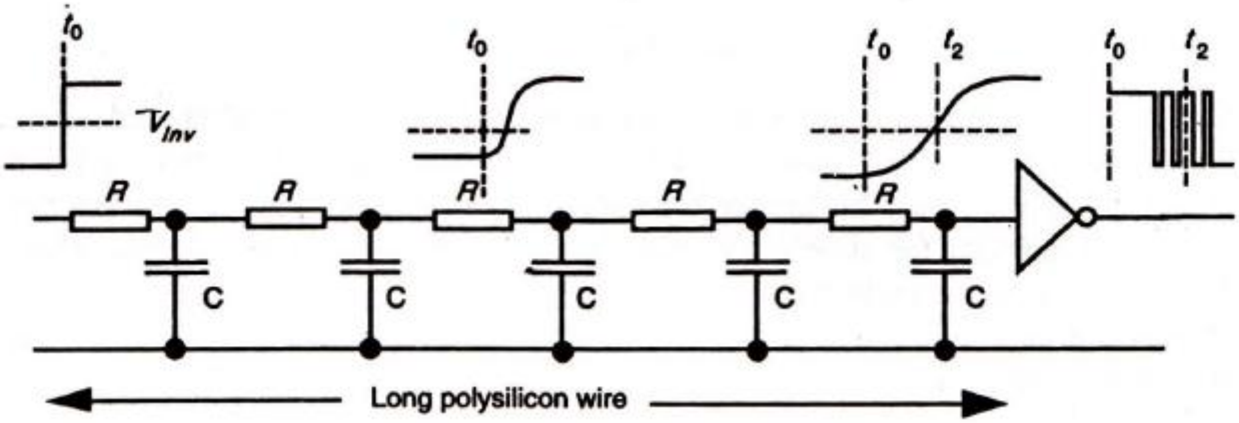
Thus, the overall delay increases rapidly as n increases and in practice no more than *four* pass transistors should be normally connected in series. However, this number can be exceeded if a buffer is inserted between each group of four pass transistors *or* if relatively long time delays are acceptable.

Design of Long Polysilicon Wires

Long polysilicon wires also contribute distributed series *R* and *C* as was the case for cascaded pass transistors and, in consequence, signal propagation is slowed down. This would also be the case for wires in diffusion where the value of *C* may be quite high, and for this reason the designer is discouraged from running signals in diffusion except over very short distances.

For long polysilicon runs, the use of buffers is recommended. In general, the use of buffers to drive long polysilicon runs has two desirable effects. First, the signal propagation is speeded up and, second, there is a reduction in sensitivity to noise.

The reason why noise may be a problem with slowly rising signals may be deduced by considering Figure 4.18. In the diagram the slow rise-time of the signal at the input of the inverter (to which the signal emerging from the long polysilicon line is connected) means that the input voltage spends a relatively long time in the vicinity of V_{inv} so that small disturbances due to noise will switch the inverter state between '0' and '1' as shown at the output point.



Note. V_{inv} = Inverter threshold

FIGURE 4.18 Possible effects of delays in long polysilicon wires.

Thus it is essential that long polysilicon wires be driven by suitable buffers to guard against the effects of noise and to speed up the rise-time of propagated signal edges.

WIRING CAPACITANCES

we considered the area capacitances associated with the layers to substrate and from gate to channel. However, there are other significant sources of capacitance which contribute to the overall wiring capacitance. Three such sources are discussed below.

Fringing Fields

Capacitance due to fringing field effects can be a major component of the overall capacitance of interconnect wires. For fine line metallization, the value of fringing field capacitance (C_{ff}) can be of the same order as that of the area capacitance. Thus, C_{ff} should be taken into account if accurate prediction of performance is needed.

$$C_{ff} = \varepsilon_{SiO_2}\varepsilon_0 l \left[\frac{\pi}{1n\left\{1 + \frac{2d}{t}\,(1 + \sqrt{(1 + \frac{t}{d}}))\right\}} - \frac{t}{4d} \right]$$

where

l = wire length

t = thickness of wire

d = wire to substrate separation

Then, total wire capacitance

$$C_w = C_{area} + C_{ff}$$

Interlayer Capacitances

Quite obviously the parallel plate effects are present between one layer and another. For example, some thought on the matter will confirm the fact that, for a given area, metal to polysilicon capacitance must be higher than metal to substrate. The reason for not taking such effects into account for simple calculations is that the effects occur only where layers cross or when one layer underlies another, and in consequence interlayer capacitance is highly dependent on layout. However, for regular structures it is readily calculated and contributes significantly to the accuracy of circuit modeling and delay calculation.

Peripheral Capacitance

The source and drain n-diffusion regions (n-active regions for Orbit processes) form junctions with the p-substrate or p-well at well-defined and uniform depths; similarly for p-diffusion (p-active) regions in n-substrates or n-wells. For diffusion regions, each diode thus formed has associated with it a peripheral (side-wall) capacitance in picofarads per unit length which, in total, can be considerably greater than the area capacitance of the diffusion region to substrate; the smaller the source or drain area, the greater becomes the relative value of the peripheral capacitance.

For Orbit processes, the n-active and p-active regions are formed by impurity implant at the surface of the silicon and thus, having negligible depth, they have negligible peripheral capacitance.

However, for n- and p-regions formed by a diffusion process, the peripheral capacitance is important and becomes particularly so as we shrink the device dimensions.

In order to calculate the total diffusion capacitance we must add the contributions of area and peripheral components

$$C_{total} = C_{area} + C_{periph}$$

CHOICE OF LAYERS

Frequently, in designing an arrangement to meet given specifications, there are several possible ways in which the requirements may be met, including the choice between the layers on which to route certain data and control signals. However, there are certain commonsense constraints which should be considered:

- V_{DD} and V_{SS} (*GND*) should be distributed on metal layers wherever possible and should not depart from metal except for 'duck unders', preferably on the diffusion layer when this is absolutely essential. A consideration of R_s values will reveal the reason for this.
- Long lengths of polysilicon should be used only after careful consideration because of the relatively high R_s value of the polysilicon layer. Polysilicon is unsuitable for routing V_{DD} or V_{SS} other than for very small distances.
- With these restrictions in mind, it is generally the case that the resistances associated with transistors are much higher than any reasonable wiring resistance, so that there is no real danger of any problem due to voltage divider effects between wiring and transistor resistances.
- Capacitive effects must also be carefully considered, particularly where fast signal lines are required and particularly in relation to signals on wiring having relatively high values of R_s. Diffusion (or active) areas have relatively high values of capacitance to substrate and are harder to drive in consequence. Charge sharing may also cause problems in certain circuits or architectures and must be carefully considered. Over small equipotential regions, the signal on a wire can be treated as being identical at all points. Within each region the delay associated with signal propagation is small in comparison with gate delays and with signal delays in systems connected by the wires.

Thus the wires in a MOS system can be modeled as simple capacitors. This concept leads to the establishment of electrical rules (guidelines) for communication paths (wires) as given in Table 4.4.

The factors set out in Tables 4.4 and 4.5 help to put matters in perspective.

TABLE 4.4 Electrical rules

Layer	*Maximum length of communication wire*		
	lambda-based (5 μm)	*μm-based (2 μm)*	*μm-based (1.2 μm)*
Metal	chip wide	chip wide	chip wide
Silicide	2,000λ	NA	NA
Polysilicon	200λ	400 μm	250 μm
Diffusion (active)	20λ*	100 μm	60 μm

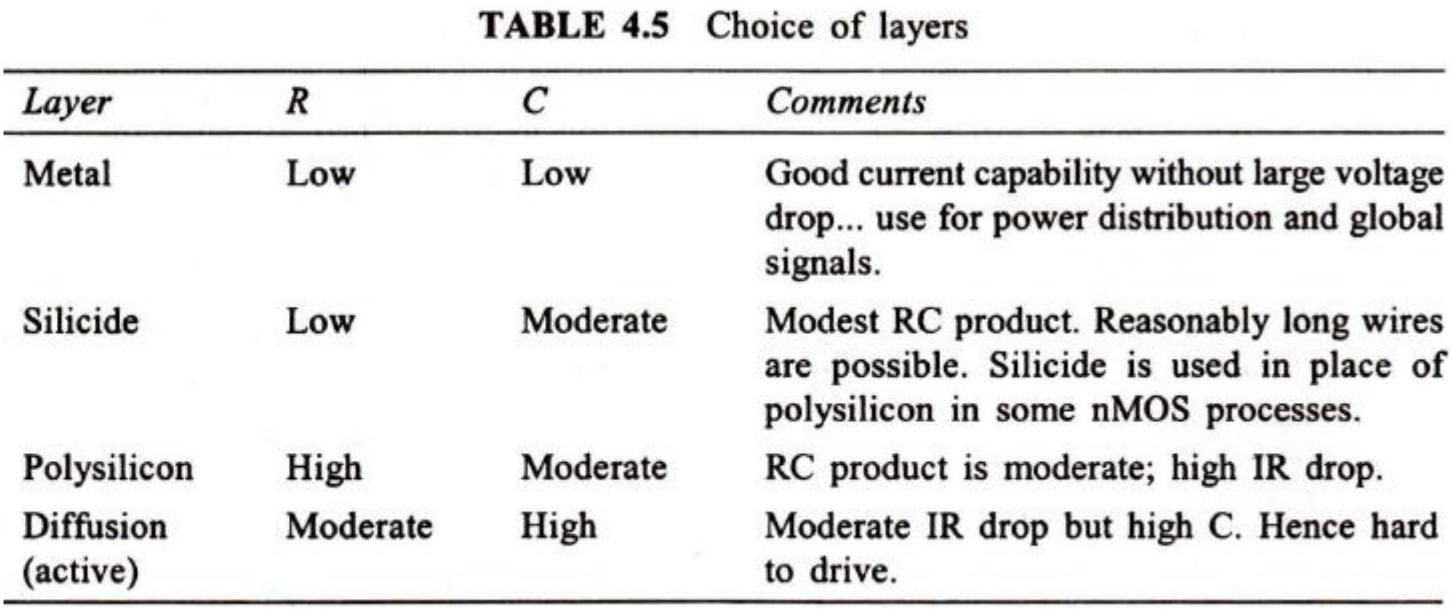
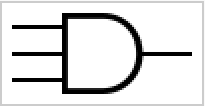


TABLE 4.5 Choice of layers

Layer	*R*	*C*	*Comments*
Metal	Low	Low	Good current capability without large voltage drop... use for power distribution and global signals.
Silicide	Low	Moderate	Modest RC product. Reasonably long wires are possible. Silicide is used in place of polysilicon in some nMOS processes.
Polysilicon	High	Moderate	RC product is moderate; high IR drop.
Diffusion (active)	Moderate	High	Moderate IR drop but high C. Hence hard to drive.

FAN IN

Fan-in is the number of inputs a gate can handle. For instance the fan-in for the AND gate shown in the figure is 3. Physical logic gates with a large fan-in tend to be slower than those with a small fan-in. This is because the complexity of the input circuitry increases the input capacitance of the device. Using logic gates with higher fan-in will help reducing the depth of a logic circuit.



An AND gate with three inputs has a fan-in of 3.

FAN OUT

In most designs, logic gates are connected to form more complex circuits. While no more than one logic gate output is connected to any single input, it is common for one output to be connected to several inputs. The technology used to implement logic gates usually allows a certain number of gate inputs to be wired directly together without additional interfacing circuitry. The maximum fan-out of an output measures its load-driving capability: it is the greatest number of inputs of gates of the same type to which the output can be safely connected.