

# A review of microelectronics and an introduction to MOS technology

*If you would have the kindness to begin at the beginning, I should be vastly obliged; all these stories that begin in the middle simply fog my wit.*

Count Anthony Hamilton

## Objectives

This chapter 'sets the scene' by reviewing the evolution of integrated circuits (ICs) and comparing the general characteristics of currently available technologies, including BiCMOS and GaAs as well as nMOS and CMOS.

Basic MOS transistor action is briefly reviewed and an overview of fabrication processes is given to help appreciate the nature of the technologies.

## 1.1 Introduction to integrated circuit technology

There is no doubt that our daily lives are significantly affected by electronic engineering technology. This is true on the domestic scene, in our professional disciplines, in the workplace, and in leisure activities. Indeed, even at school, tomorrow's adults are exposed to and are coming to terms with quite sophisticated electronic devices and systems. There is no doubt that revolutionary changes have taken place in a relatively short time and it is also certain that even more dramatic advances will be made in the next decade.

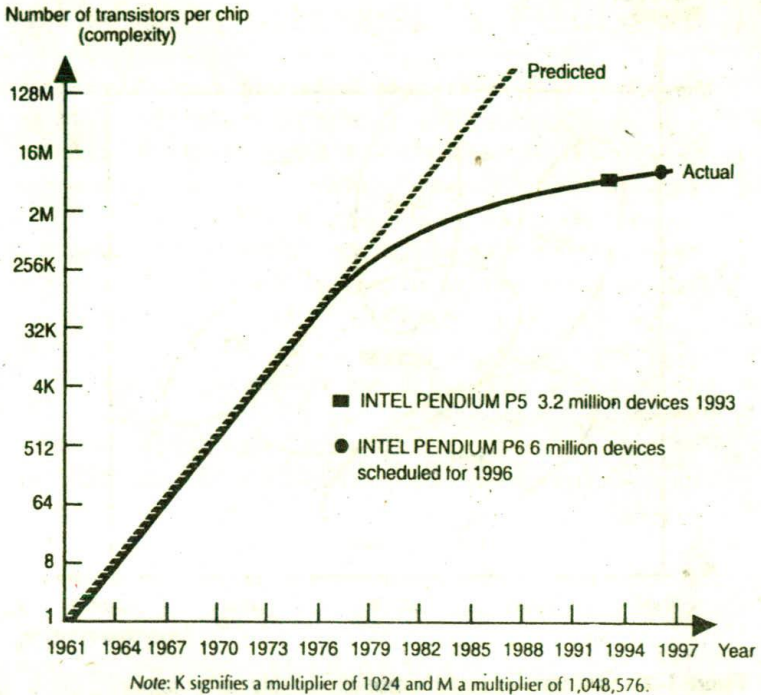
Electronics as we know it today is characterized by reliability, low power dissipation, extremely low weight and volume, and low cost, coupled with an ability to cope easily with a high degree of sophistication and complexity. Electronics, and in particular the integrated circuit, has made possible the design of powerful and flexible processors which provide highly intelligent and adaptable devices for the user. Integrated circuit memories have provided the essential elements to complement these processors and, together with a wide range of logic and analog integrated circuitry, they have provided the system designer with components of considerable capability and extensive application. Furthermore, the revolutionary advances in technology have not yet by any means run their full course and the potential for future developments is exciting to say the least.

Up until the 1950s, electronic active device technology was dominated by the vacuum tube and, although a measure of miniaturization and circuit integration did take place, the technology did not lend itself to miniaturization as we have come to accept it today. Thus the vast majority of present-day electronics is the result of the invention of the transistor in 1947.

The invention of the transistor by William B. Shockley, Walter H. Brattain and John Bardeen of Bell Telephone Laboratories was followed by the development of the Integrated Circuit (IC). The very first IC emerged at the beginning of 1960 and since that time there have already been four generations of ICs: SSI (small scale integration), MSI (medium scale integration), LSI (large scale integration), and VLSI (very large scale integration). Now we are beginning to see the emergence of the fifth generation, ULSI (ultra large scale integration), which is characterized by complexities in excess of 3 million devices on a single IC chip. Further miniaturization is still to come and more revolutionary advances in the application of this technology must inevitably occur.

Over the past several years, Silicon CMOS technology has become the dominant fabrication process for relatively high performance and cost effective VLSI circuits. The revolutionary nature of this development is indicated by the way in which the number of transistors integrated in circuits on a single chip has grown, as indicated in Figure 1-1. Such progress is highlighted by recent products such as RISC chips in which it is possible to process some 35 million instructions per second. In order to improve on this throughput rate it will be necessary to improve





**Figure 1-1** Moore's first law: transistors integrated on a single chip (commercial products)

the technology, both in terms of scaling and processing, and through the incorporation of other enhancements such as BiCMOS. The implication of this approach is that existing silicon technology could effectively facilitate the tripling of rate. Beyond this, that is, above 100 million instructions per second, one must look to other technologies. In particular, the emerging gallium arsenide (GaAs) based technology will be most significant in this area of ultra high speed logic/fast digital processors. GaAs also has further potential as a result of its photo-electronic properties, both as a receiver and as a transmitter of light. GaAs in combination with silicon will provide the designer with some very exciting possibilities.

It is most informative in assessing the role of the currently available technologies to review their speed and power performance domains. This has been set out as Figure 1-2 and the potential presented by each may be readily assessed.

This text deals mostly with silicon-based VLSI, including BiCMOS, but also introduces GaAs-based technology. ECL-based technology is not covered here, but much of the material given is relevant to the general area of the design of digital integrated circuits.

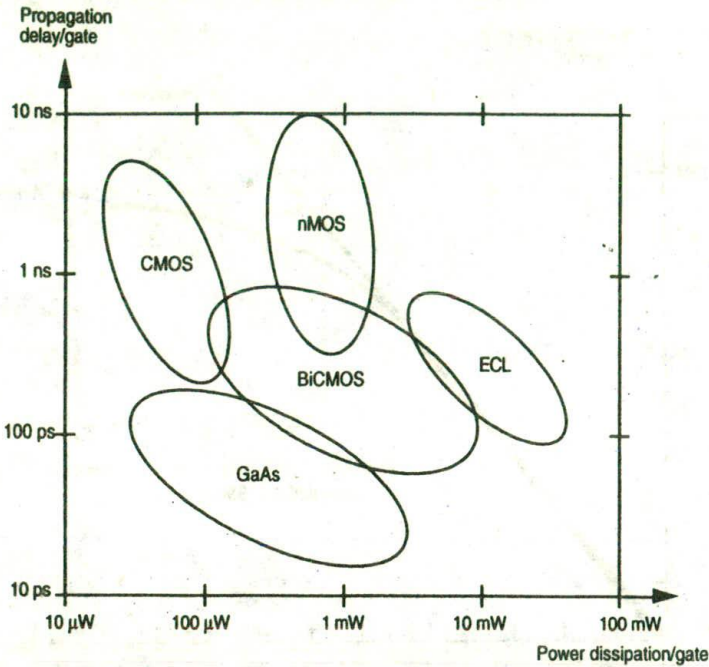


Figure 1-2 Speed/power performance of available technologies

## 1.2 The integrated circuit (IC) era

Such has been the potential of the silicon integrated circuit that there has been an extremely rapid growth in the number of transistors (as a measure of complexity) being integrated into circuits on a single silicon chip. In less than three decades, this number has risen from tens to millions as can be seen in Figure 1-1. The figure sets out what has become known as 'Moore's first law' after predictions made by Gordon Moore (of Intel) in the 1960s. It may be seen that his predictions have largely come true except for an increasing divergence between 'predicted' and 'actual' over the last few years due to problems associated with the complexities involved in designing and testing such very large circuits.

Such has been the impact of this revolutionary growth that IC technology now affects almost every aspect of our lives. More is still to come since we have not yet reached the limits of miniaturization and there is no doubt that tens of millions of transistors will be readily integrated onto a single chip in the future. This evolutionary process is reflected in Table 1-1.

Truly the 1970s, the 1980s and now the 1990s may well be described as the integrated circuit era.



Table 1-1 Microelectronics evolution

| Year   | 1947 | 1950                        | 1961                          | 1966  | 1971                               | 1980                                 | 1990  | 2000  |             |
|--|------|-----------------------------|-------------------------------|---|------------------------------------|--------------------------------------|---|---|-------------|
| Technology   |      | Invention of the transistor | Discrete components           | SSI   | MSI                                | LSI                                  | VLSI  | ULSI*   | GST†        |
| Approximate numbers of transistors per chip in commercial products |      | 1                           | 1                             | 10  | 100-1000                           | 1000-20,000                          | 20,000-1,000,000  | 1,000,000-10,000,000  | >10,000,000 |
| Typical products   |      | —                           | Junction Transistor and diode | Planar devices<br>Logic gates<br>Flip-flops | Counters<br>Multiplexers<br>Adders | 8 bit micro-processors<br>ROM<br>RAM | 16 and 32 bit micro-processors<br>Sophisticated peripherals<br>GHM Dram | Special processors,<br>Virtual reality machines,<br>smart sensors |             |

\* Ultra large-scale integration  
† Giant-scale integration

Note: The boundary lines between technologies in the table are not artificially created. Crossing each boundary requires new design methodology, simulation approaches, and new methods for determining and routing communications and for handling complexity.

## 1.3 Metal-oxide-semiconductor (MOS) and related VLSI technology

Within the bounds of MOS technology, the possible circuit realizations may be based on pMOS, nMOS, CMOS and now BiCMOS devices.

However, this text will deal with nMOS, then with CMOS (which includes nMOS and pMOS transistors) and BiCMOS, and finally with GaAs technology, all of which may be classed as leading integrated circuit technologies.

Although CMOS is the dominant technology, some of the examples used to illustrate the design processes will be presented in nMOS form. The reasons for this are as follows:

- For nMOS technology, the design methodology and the design rules are easily learned, thus providing a simple but excellent introduction to structured design for VLSI.
- nMOS technology and design processes provide an excellent background for other technologies. In particular, some familiarity with nMOS allows a relatively easy transition to CMOS technology and design.
- For GaAs technology some arrangements in relation to logic design are similar to those employed in nMOS technology. Therefore, understanding the basics of nMOS design will assist in the layout of GaAs circuits.

Not only is VLSI technology providing the user with a new and more complex range of 'off the shelf' circuits, but VLSI design processes are such that system designers can readily design their own special circuits of considerable complexity. This provides a new degree of freedom for designers and it is probable that some very significant advances will result. Couple this with the fact that integration density is increasing rapidly, as advances in technology shrink the feature size for circuits integrated in silicon. Typical manufacturers' commercial IC products have shown this trend quite clearly as indicated in Figure 1-3 and, simultaneously, the effectiveness of the circuits produced has increased with scaling down. A common measure of effectiveness is the speed power product of the basic logic gate circuit of the technology (for nMOS, the *Nor* gate; with *Nand* and *Nor* gates for CMOS). Speed power product is measured in picojoules (pJ) and is the product of the gate switching delay in nanoseconds and the gate power dissipation in milliwatts. Typical figures are implied in Figure 1-2.

## 1.4 Basic MOS transistors

Having now established some background, let us turn our attention to basic MOS processes and devices. In particular, let us examine the basic nMOS enhancement and depletion mode transistors as shown in Figures 1-4 (a) and (b).



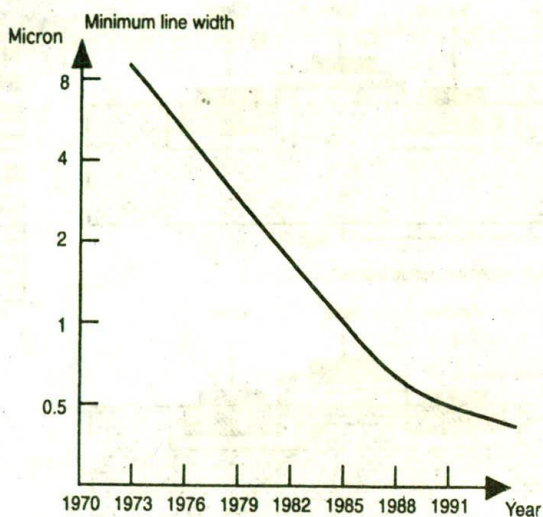


Figure 1-3 Approximate minimum line width of commercial products versus year

nMOS devices are formed in a p-type substrate of moderate doping level. The source and drain regions are formed by diffusing n-type impurities through suitable masks into these areas to give the desired n-impurity concentration and give rise to depletion regions which extend mainly in the more lightly doped p-region as shown. Thus, source and drain are isolated from one another by two diodes. Connections to the source and drain are made by a deposited metal layer. In order to make a useful device, there must be the capability for establishing and controlling a current between source and drain, and this is commonly achieved in one of two ways, giving rise to the enhancement mode and depletion mode transistors.

Consider the enhancement mode device first, shown in Figure 1-4(a). A polysilicon gate is deposited on a layer of insulation over the region between source and drain. Figure 1-4(a) shows a basic enhancement mode device in which the channel is not established and the device is in a non-conducting condition,  $V_D = V_S = V_{gs} = 0$ . If this gate is connected to a suitable positive voltage with respect to the source, then the electric field established between the gate and the substrate gives rise to a charge inversion region in the substrate under the gate insulation and a conducting path or channel is formed between source and drain.

The channel may also be established so that it is present under the condition  $V_{gs} = 0$  by implanting suitable impurities in the region between source and drain during manufacture and prior to depositing the insulation and the gate. This arrangement is shown in Figure 1-4(b). Under these circumstances, source and drain are connected by a conducting channel, but the channel may now be closed by applying a suitable negative voltage to the gate.

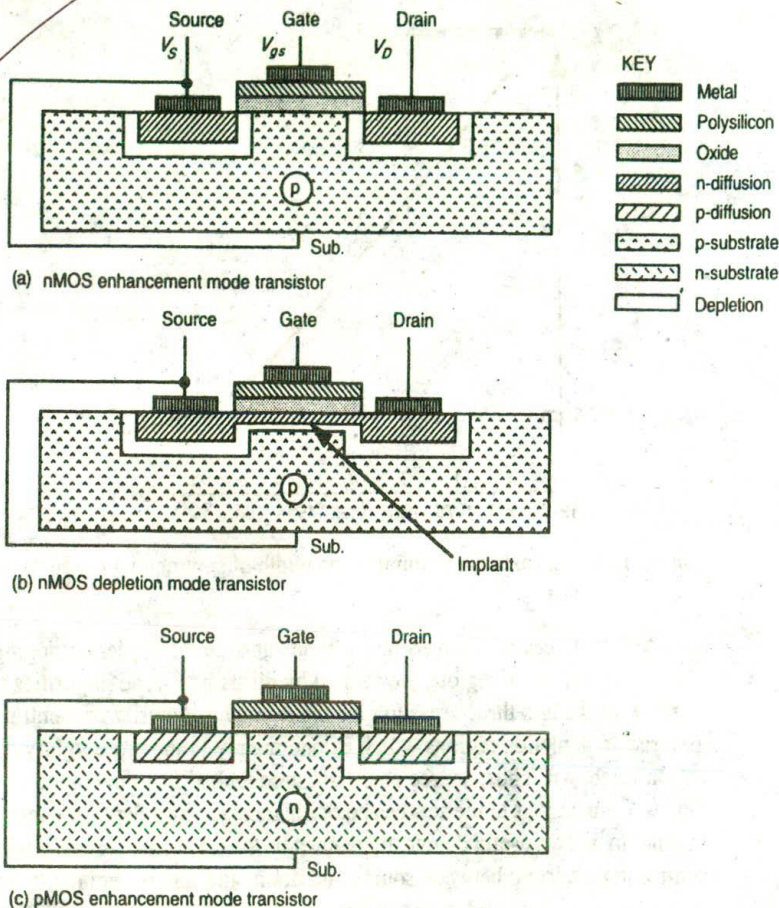


Figure 1-4 MOS transistors ( $V_D = 0V$ . Source gate and substrate to 0V)

In both cases, variations of the gate voltage allow control of any current flow between source and drain.

Figure 1-4(c) shows the basic pMOS transistor structure for an enhancement mode device. In this case the substrate is of n-type material and the source and drain diffusions are consequently p-type. In the figure, the conditions shown are those for an unbiased device; however, the application of a *negative* voltage of suitable magnitude ( $> |V_t|$ ) between gate and source will give rise to the formation of a channel (p-type) between the source and drain and current may then flow if the drain is made negative with respect to the source. In this case the current is carried by holes as opposed to electrons (as is the case for nMOS devices). In consequence, pMOS transistors are inherently slower than nMOS, since hole mobility  $\mu_p$  is less, by a factor of approximately 2.5, than electron mobility  $\mu_n$ .



However, bearing these differences in mind, the discussions of nMOS transistors which follow relate equally well to pMOS transistors.

## 1.5 Enhancement mode transistor action

To gain some understanding of this mechanism, let us further consider the enhancement mode device, as in Figure 1-5, under three sets of conditions. It must first be recognized that in order to establish the channel in the first place, a minimum voltage level of *threshold voltage*  $V_t$  must be established between gate

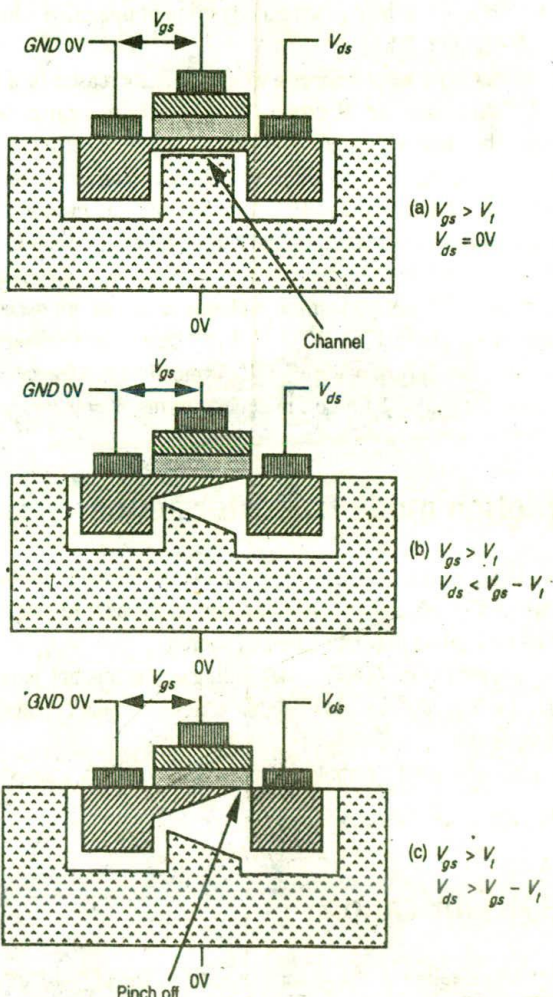


Figure 1-5 Enhancement mode transistor for particular values of  $V_{ds}$  with ( $V_{gs} > V_t$ )

and source (and of course between gate and substrate as a result). Figure 1-5(a) then indicates the conditions prevailing with the channel established but no current flowing between source and drain ( $V_{ds} = 0$ ). Now consider the conditions prevailing when current flows in the channel by applying a voltage  $V_{ds}$  between drain and source. There must, of course, be a corresponding IR drop  $= V_{ds}$  along the channel. This results in the voltage between gate and channel varying with distance along the channel with the voltage being a maximum of  $V_{gs}$  at the source end. Since the effective gate voltage is  $V_g = V_{gs} - V_t$  (no current flows when  $V_{gs} < V_t$ ), there will be voltage available to invert the channel at the drain end so long as  $V_{gs} - V_t \geq V_{ds}$ . The limiting condition comes when  $V_{ds} = V_{gs} - V_t$ . For all voltages  $V_{ds} < V_{gs} - V_t$ , the device is in the non-saturated region of operation which is the condition shown in Figure 1-5(b).

Consider now what happens when  $V_{ds}$  is increased to a level greater than  $V_{gs} - V_t$ . In this case, an IR drop  $= V_{gs} - V_t$  takes place over less than the whole length of the channel so that over part of the channel, near the drain, there is insufficient electric field available to give rise to an inversion layer to create the channel. The channel is therefore 'pinched off', as indicated in Figure 1-5(c). Diffusion current completes the path from source to drain in this case, causing the channel to exhibit a high resistance and behave as a constant current source. This region, known as *saturation*, is characterized by almost constant current for increase of  $V_{ds}$  above  $V_{ds} = V_{gs} - V_t$ . In all cases, the channel will cease to exist and no current will flow when  $V_{gs} < V_t$ . Typically, for enhancement mode devices,  $V_t = 1$  volt for  $V_{DD} = 5$  volt or, in general terms,  $V_t = 0.2 V_{DD}$ .

## 1.6 Depletion mode transistor action

For depletion mode devices the channel is established, because of the implant, even when  $V_{gs} = 0$ , and to cause the channel to cease to exist a negative voltage  $V_{td}$  must be applied between gate and source.

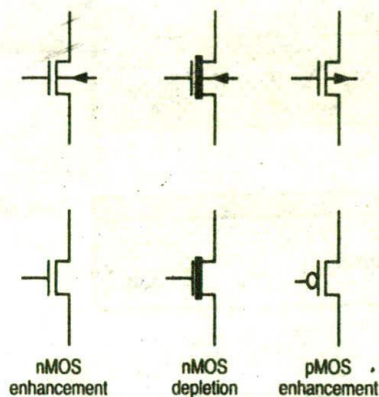
$V_{td}$  is typically  $< -0.8 V_{DD}$ , depending on the implant and substrate bias, but, threshold voltage differences aside, the action is similar to that of the enhancement mode transistor.

Commonly used symbols for nMOS and pMOS transistors are set out in Figure 1-6.

## 1.7 nMOS fabrication

A brief introduction to the general aspects of the polysilicon gate self-aligning nMOS fabrication process will now be given. As well as being relevant in their own right, the fabrication processes used for nMOS are relevant to CMOS and





**Figure 1-6** Transistor circuit symbols

BiCMOS which may be viewed as involving additional fabrication steps. Also, it is clear that an appreciation of the fabrication processes will give an insight into the way in which design information must be presented and into the reasons for certain performance characteristics and limitations. An nMOS process is illustrated in Figure 1-7 and may be outlined as follows:

1. Processing is carried out on a thin wafer cut from a single crystal of silicon of high purity into which the required p-impurities are introduced as the crystal is grown. Such wafers are typically 75 to 150 mm in diameter and 0.4 mm thick and are doped with, say, boron to impurity concentrations of  $10^{15}/\text{cm}^3$  to  $10^{16}/\text{cm}^3$ , giving resistivity in the approximate range 25 ohm cm to 2 ohm cm.
2. A layer of silicon dioxide ( $\text{SiO}_2$ ), typically 1  $\mu\text{m}$  thick, is grown all over the surface of the wafer to protect the surface, act as a barrier to dopants during processing, and provide a generally insulating substrate onto which other layers may be deposited and patterned.
3. The surface is now covered with a photoresist which is deposited onto the wafer and spun to achieve an even distribution of the required thickness.
4. The photoresist layer is then exposed to ultraviolet light through a mask which defines those regions into which diffusion is to take place together with transistor channels. Assume, for example, that those areas exposed to ultraviolet radiation are polymerized (hardened), but that the areas required for diffusion are shielded by the mask and remain unaffected.
5. These areas are subsequently readily etched away together with the underlying silicon dioxide so that the wafer surface is exposed in the window defined by the mask.

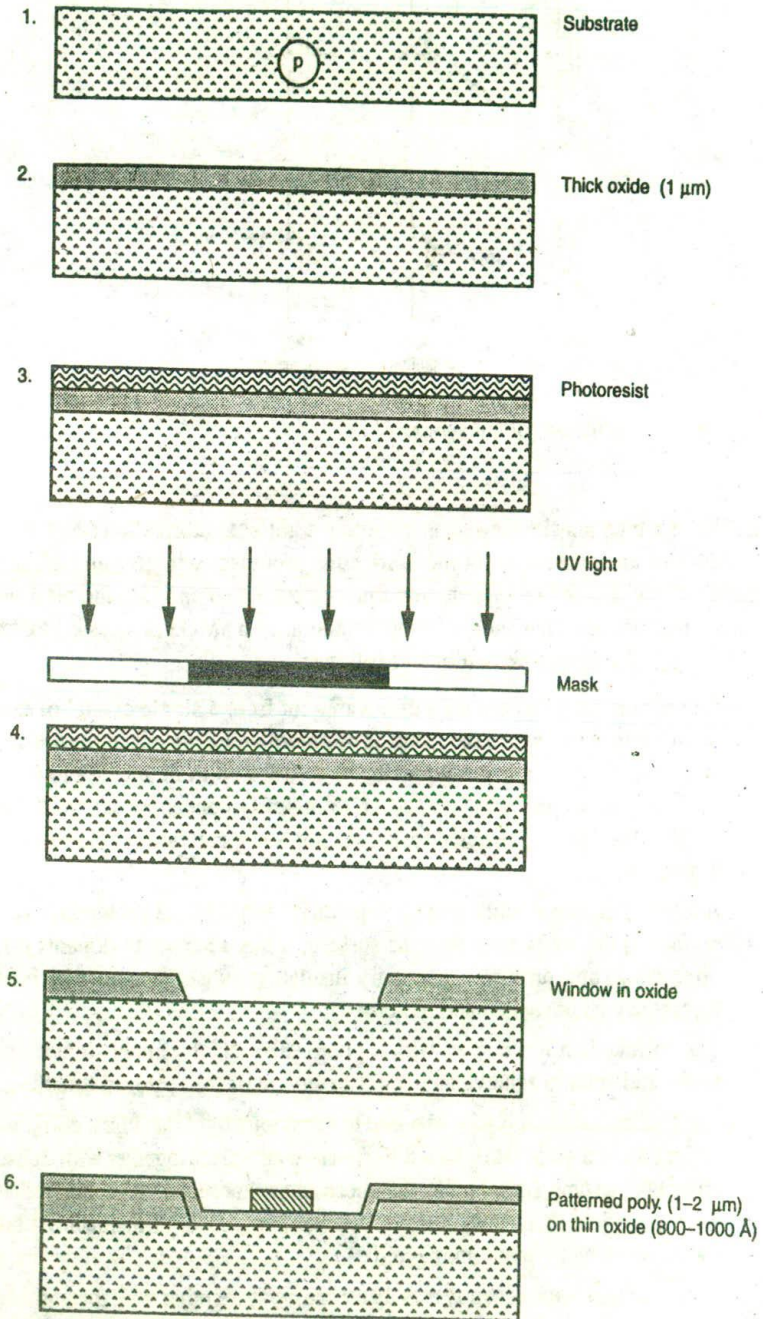


Figure 1-7 nMOS fabrication process



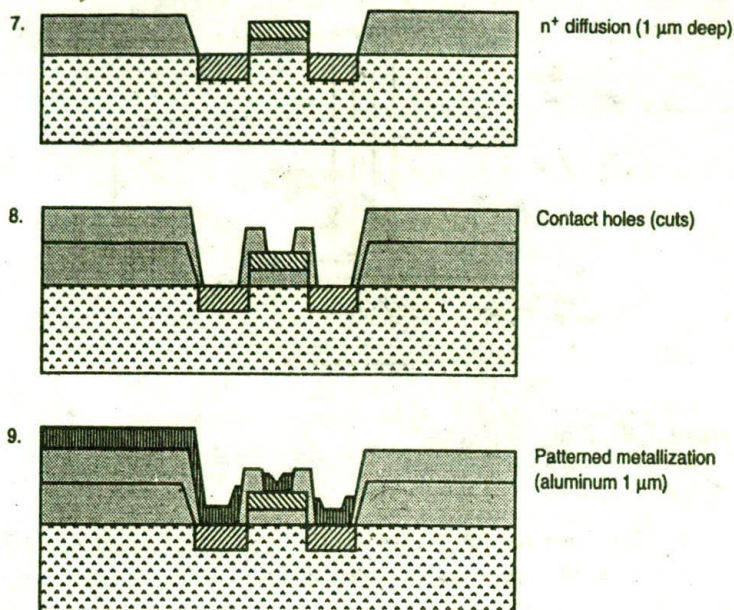


Figure 1-7 continued

6. The remaining photoresist is removed and a thin layer of  $\text{SiO}_2$  (0.1  $\mu\text{m}$  typical) is grown over the entire chip surface and then polysilicon is deposited on top of this to form the gate structure. The polysilicon layer consists of heavily doped polysilicon deposited by chemical vapor deposition (CVD). In the fabrication of fine pattern devices, precise control of thickness, impurity concentration, and resistivity is necessary.
7. Further photoresist coating and masking allows the polysilicon to be patterned (as shown in Step 6), and then the thin oxide is removed to expose areas into which n-type impurities are to be diffused to form the source and drain as shown. Diffusion is achieved by heating the wafer to a high temperature and passing a gas containing the desired n-type impurity (for example, phosphorus) over the surface as indicated in Figure 1-8. Note that the polysilicon with underlying thin oxide and the thick oxide act as masks during diffusion — the process is self-aligning.
8. Thick oxide ( $\text{SiO}_2$ ) is grown over all again and is then masked with photoresist and etched to expose selected areas of the polysilicon gate and the drain and source areas where connections (i.e. contact cuts) are to be made.
9. The whole chip then has metal (aluminum) deposited over its surface to a thickness typically of 1  $\mu\text{m}$ . This metal layer is then masked and etched to form the required interconnection pattern.

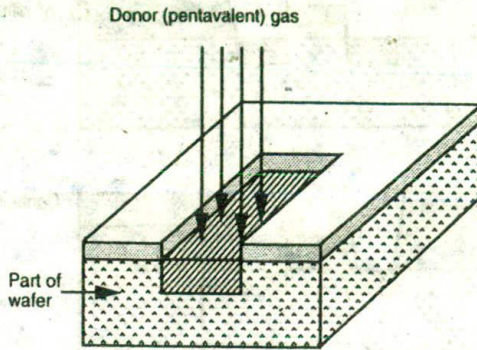


Figure 1-8 Diffusion process

It will be seen that the process revolves around the formation or deposition and patterning of three layers, separated by silicon dioxide insulation. The layers are diffusion within the substrate, polysilicon on oxide on the substrate, and metal insulated again by oxide.

To form depletion mode devices it is only necessary to introduce a masked ion implantation step between Steps 5 and 6 in Figure 1-7. Again, the thick oxide acts as a mask and this process stage is also self-aligning.

Consideration of the processing steps will reveal that relatively few masks are needed and the self-aligning aspects of the masking processes greatly ease the problems of mask registration. In practice, some extra process steps are necessary, including the overglassing of the whole wafer, except where contacts to the outside world are required. However, the process is basically straightforward to envisage and circuit design eventually comes down to the business of delineating the masks for each stage of the process. The essence of the process may be reiterated as follows.

### 1.7.1 Summary of an nMOS process

- Processing takes place on a p-doped silicon crystal wafer on which is grown a 'thick' layer of  $\text{SiO}_2$ .
- *Mask 1* — Pattern  $\text{SiO}_2$  to expose the silicon surface in areas where paths in the diffusion layer or source, drain or gate areas of transistors are required. Deposit thin oxide over all. For this reason, this mask is often known as the 'thinox' mask but some texts refer to it as the *diffusion mask*.
- *Mask 2* — Pattern the ion implantation within the thinox region where depletion mode devices are to be produced — *self-aligning*.



- *Mask 3* — Deposit polysilicon over all (1.5  $\mu\text{m}$  thick typically), then pattern using Mask 3. Using the same mask, remove thin oxide layer where it is not covered by polysilicon.
- Diffuse  $n^+$  regions into areas where thin oxide has been removed. Transistor drains and sources are thus self-aligning with respect to the gate structures.
- *Mask 4* — Grow thick oxide over all and then etch for contact cuts.
- *Mask 5* — Deposit metal and pattern with Mask 5.
- *Mask 6* — Would be required for the overglassing process step.

## 1.8 CMOS fabrication

There are a number of approaches to CMOS fabrication, including the p-well, the n-well, the twin-tub, and the silicon-on-insulator processes. In order to introduce the reader to CMOS design we will be concerned mainly with well-based circuits. The p-well process is widely used in practice and the n-well process is also popular, particularly as it was an easy retrofit to existing nMOS lines.

For the lambda-based rules set out later, we will assume a p-well process.

### 1.8.1 The p-well process

A brief overview of the fabrication steps may be obtained with reference to Figure 1-9, noting that the basic processing steps are of the same nature as those used for nMOS.

In primitive terms, the structure consists of an n-type substrate in which p-devices may be formed by suitable masking and diffusion and, in order to accommodate n-type devices, a deep p-well is diffused into the n-type substrate as shown.

This diffusion must be carried out with special care since the p-well doping concentration and depth will affect the threshold voltages as well as the breakdown voltages of the n-transistors. To achieve low threshold voltages (0.6 to 1.0 V), we need either deep well diffusion or high well resistivity. However, deep wells require larger spacing between the n- and p-type transistors and wires because of lateral diffusion and therefore a larger chip area.

The p-wells act as substrates for the n-devices within the parent n-substrate, and, provided that voltage polarity restrictions are observed, the two areas are electrically isolated. However, since there are now in effect two substrates, two substrate connections ( $V_{DD}$  and  $V_{SS}$ ) are required, as shown in Figure 1-10.

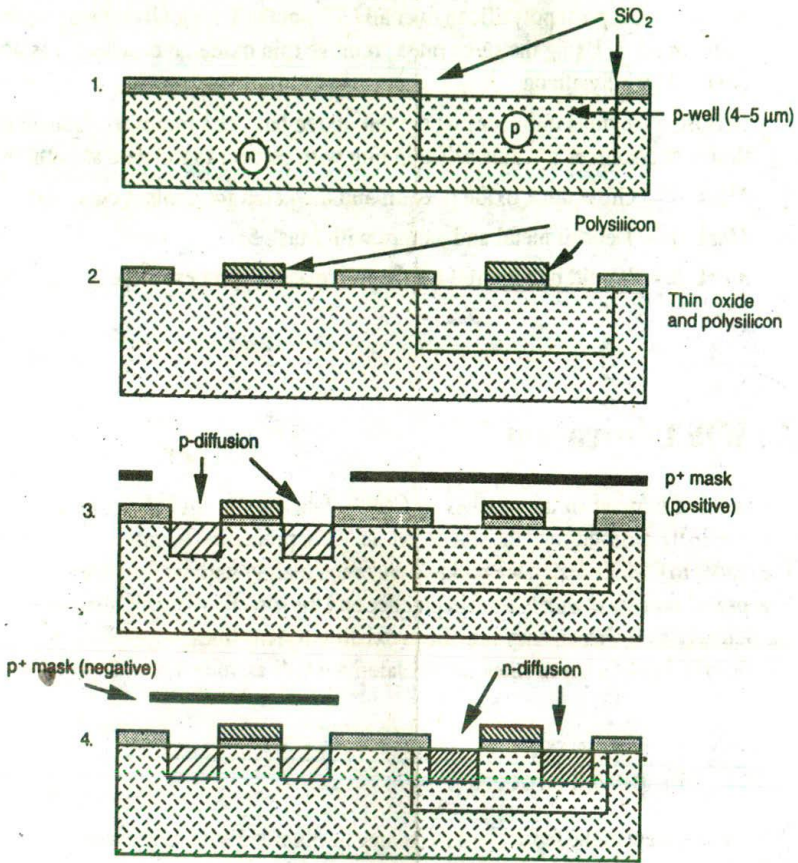


Figure 1-9 CMOS p-well process steps

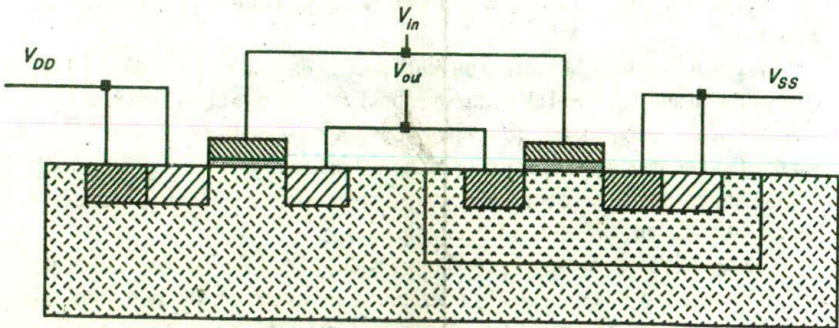


Figure 1-10 CMOS p-well inverter showing V<sub>DD</sub> and V<sub>SS</sub> substrate connections



In all other respects — masking, patterning, and diffusion — the process is similar to nMOS fabrication. In summary, typical processing steps are:

- *Mask 1* — defines the areas in which the deep p-well diffusions are to take place.
- *Mask 2* — defines the thinox regions, namely those areas where the thick oxide is to be stripped and thin oxide grown to accommodate p- and n-transistors and diffusion wires.
- *Mask 3* — used to pattern the polysilicon layer which is deposited after the thin oxide.
- *Mask 4* — A p-plus mask is now used (to be in effect 'Anded' with Mask 2) to define all areas where p-diffusion is to take place.
- *Mask 5* — This is usually performed using the negative form of the p-plus mask and, with Mask 2, defines those areas where n-type diffusion is to take place.
- *Mask 6* — Contact cuts are now defined.
- *Mask 7* — The metal layer pattern is defined by this mask.
- *Mask 8* — An overall passivation (overglass) layer is now applied and Mask 8 is needed to define the openings for access to bonding pads.

### 1.8.2 The n-well process

As indicated earlier, although the p-well process is widely used, n-well fabrication has also gained wide acceptance, initially as a retrofit to nMOS lines.

N-well CMOS circuits are also superior to p-well because of the lower substrate bias effects on transistor threshold voltage and inherently lower parasitic capacitances associated with source and drain regions.

Typical n-well fabrication steps are illustrated in Figure 1-11. The first mask defines the n-well regions. This is followed by a low dose phosphorus implant driven in by a high temperature diffusion step to form the n-wells. The well depth is optimized to ensure against p-substrate to p<sup>+</sup> diffusion breakdown without compromising the n-well to n<sup>+</sup> mask separation. The next steps are to define the devices and diffusion paths, grow field oxide, deposit and pattern the polysilicon, carry out the diffusions, make contact cuts, and finally metallize as before.

It will be seen that an n<sup>+</sup> mask and its complement may be used to define the n- and p-diffusion regions respectively. These same masks also include the V<sub>DD</sub> and V<sub>SS</sub> contacts (respectively). It should be noted that, alternatively, we could have used a p<sup>+</sup> mask and its complement, since the n<sup>+</sup> and p<sup>+</sup> masks are generally complementary.

By way of illustration, Figure 1-12 shows an inverter circuit fabricated by the n-well process, and this may be directly compared with Figure 1-10.

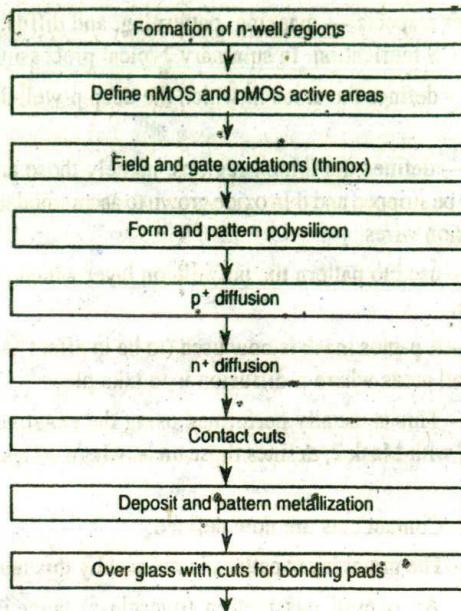


Figure 1-11 Main steps in a typical n-well process

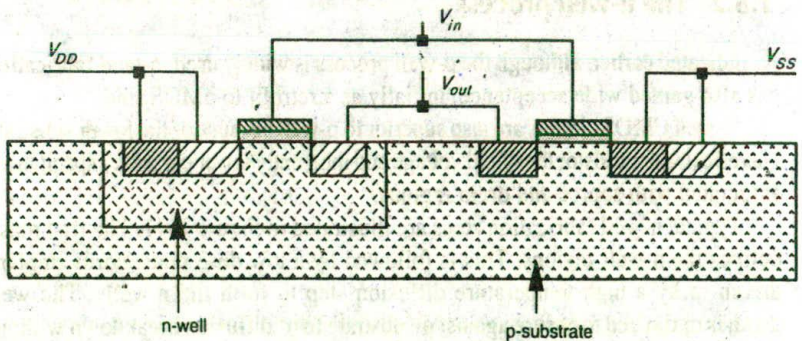


Figure 1-12 Cross-sectional view of n-well CMOS inverter

Owing to differences in charge carrier mobilities, the n-well process creates non-optimum p-channel characteristics. However, in many CMOS designs (such as domino-logic and dynamic-logic structures), this is relatively unimportant since they contain a preponderance of n-channel devices. Thus the n-channel transistors are mainly those used to form logic elements, providing speed and high density of elements.



Latch-up problems can be considerably reduced by using a low-resistivity epitaxial p-type substrate as the starting material, which can subsequently act as a very low resistance ground-plane to collect substrate currents.

However, a factor of the n-well process is that the performance of the already poorly performing p-transistor is even further degraded. Modern process lines have come to grips with these problems, and good device performance may be achieved for both p-well and n-well fabrication.

The design rules which are presented for 1.2  $\mu\text{m}$  and 2  $\mu\text{m}$  technologies in this text are for Orbif<sup>TM</sup> n-well processes.

### 1.8.2.1 The Berkeley n-well process

There are a number of p-well and n-well fabrication processes, and, in order to look more closely at typical fabrication steps, we will use the Berkeley n-well process as an example. This process is illustrated in Figure 1-13.

## 1.8.3 The twin-tub process

A logical extension of the p-well and n-well approaches is the twin-tub fabrication process.

Here we start with a substrate of high resistivity n-type material and then create both n-well and p-well regions. Through this process it is possible to preserve the performance of n-transistors without compromising the p-transistors. Doping control is more readily achieved and some relaxation in manufacturing tolerances results. This is particularly important as far as latch-up is concerned.

In general, the twin-tub process allows separate optimization of the n- and p-transistors. The arrangement of an inverter is illustrated in Figure 1-14, which may in turn be compared with Figures 1-10 and 1-12.

## 1.9 Thermal aspects of processing

The processes involved in making nMOS and CMOS devices have differing high temperature sequences as indicated in Figure 1-15.

The CMOS p-well process, for example, has a high temperature p-well diffusion process (1100 to 1250°C), the nMOS process having no such requirement. Because of the simplicity, ease of fabrication, and high density per unit area of nMOS circuits, many of the earlier IC designs, still in current use, have been fabricated using nMOS technology, and it is likely that nMOS and CMOS system designs will continue to coexist for some time to come.

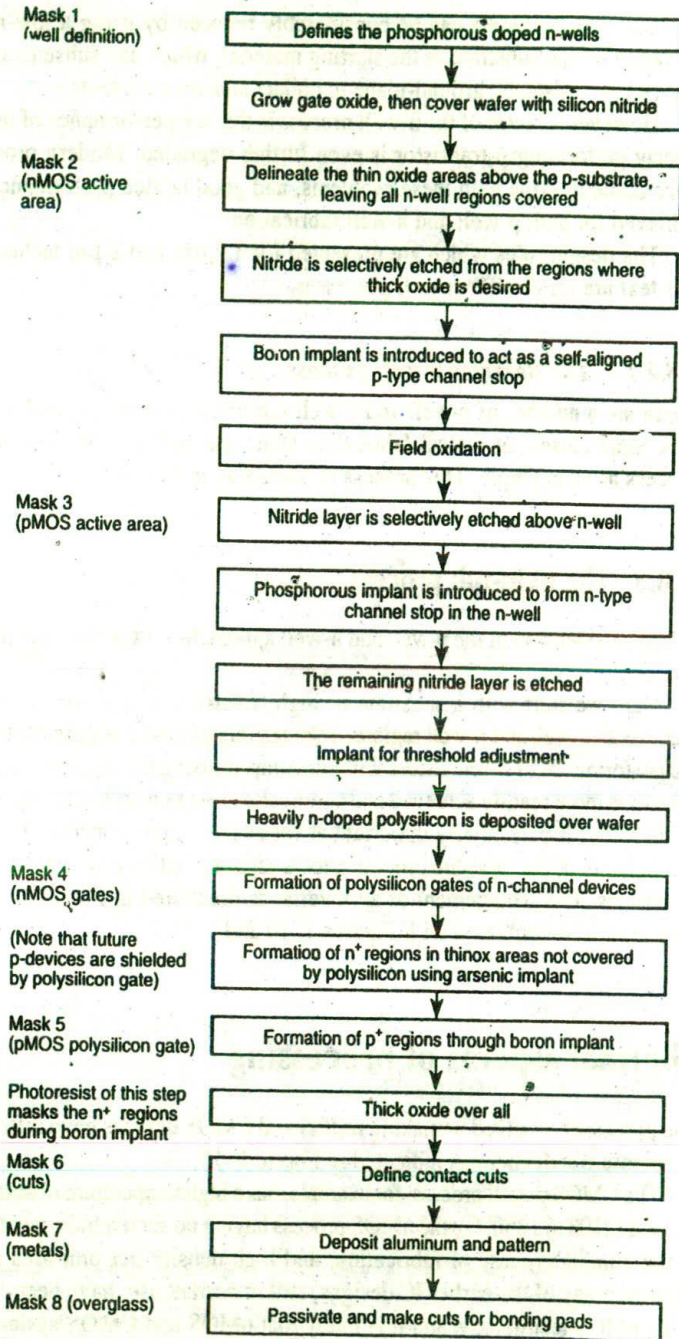


Figure 1-13 Flow diagram of Berkeley n-well fabrication



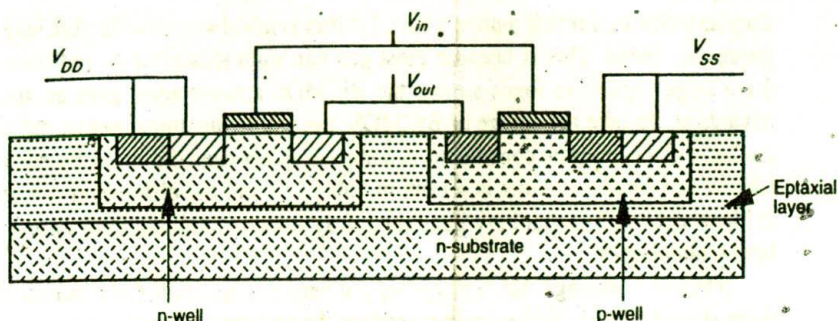


Figure 1-14 Twin-tub structure

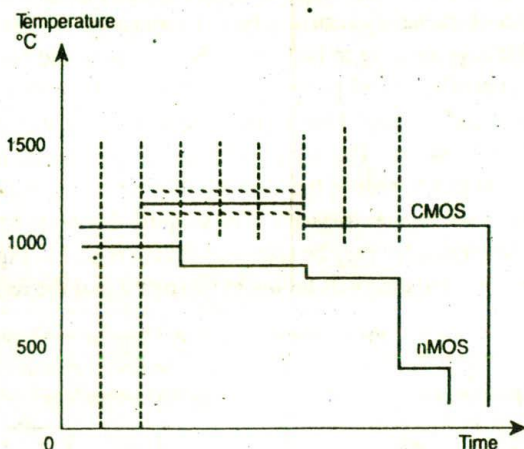


Figure 1-15 Thermal sequence difference between nMOS and CMOS processes

## 1.10 BiCMOS technology

A known deficiency of MOS technology lies in the limited load driving capabilities of MOS transistors. This is due to the limited current sourcing and current sinking abilities associated with both p- and n-transistors and although it is possible, for example, to design so-called super buffers using MOS transistors alone, such arrangements do not always compare well with the capabilities of bipolar transistors. Bipolar transistors also provide higher gain and have generally better noise and high frequency characteristics than MOS transistors and it may be seen (Figure 1-2) that using BiCMOS gates may be an effective way of speeding up VLSI circuits. However, the application of BiCMOS in subsystems such as ALU, ROM,

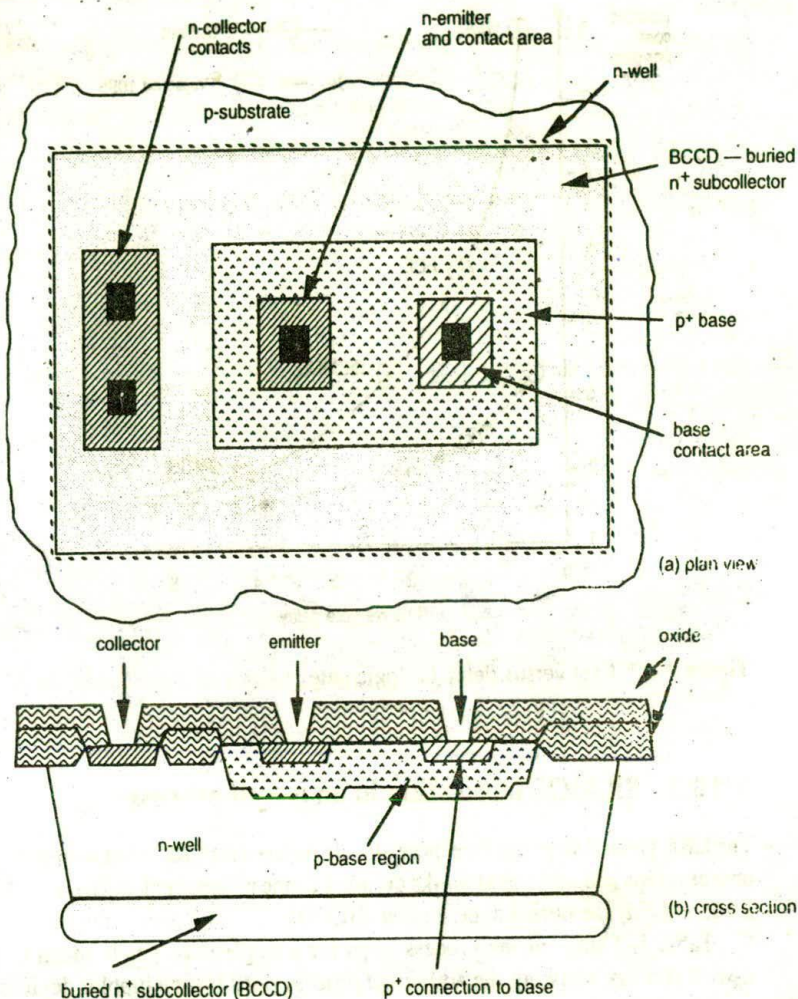
a register-file, or, for that matter, a barrel shifter is not always an effective way of improving speed. This is because most gates in such structures do not have to drive large capacitive loads so that the BiCMOS arrangements give no speed advantage. To take advantage of BiCMOS, the whole functional entity, not just the logic gates, must be considered. A comparison between the characteristics of CMOS and bipolar circuits is set out in Table 1-2 and the differences are self-evident. BiCMOS technology goes some way toward combining the virtues of both technologies.

When considering CMOS technology, it becomes apparent that theoretically there should be little difficulty in extending the fabrication processes to include bipolar as well as MOS transistors. Indeed, a problem of p-well and n-well CMOS processing is that parasitic bipolar transistors are inadvertently formed as part of the outcome of fabrication. The production of npn bipolar transistors with good performance characteristics can be achieved, for example, by extending the standard n-well CMOS processing to include further masks to add two additional layers — the  $n^+$  subcollector and  $p^+$  base layers. The npn transistor is formed in an n-well and the additional  $p^+$  base region is located in the well to form the p-base region of the transistor. The second additional layer, the buried  $n^+$  subcollector (BCCD), is added to reduce the n-well (collector) resistance and thus improve the quality of the bipolar transistor. The simplified general arrangement of such a bipolar npn transistor may be appreciated with regard to Figure 1-16. Bipolar transistor characteristics will follow in Chapter 2 and the relevant design rules

**Table 1-2** Comparisons between CMOS and bipolar technologies

| <i>CMOS technology</i>  | <i>Bipolar technology</i>  |
|---|--|
| <ul style="list-style-type: none"> <li>• Low static power dissipation</li> <li>• High input impedance (low drive current)</li> <li>• Scalable threshold voltage</li> <li>• High noise margin</li> <li>• High packing density</li> <li>• High delay sensitivity to load (fan-out limitations)</li> <li>• Low output drive current</li> <li>• Low <math>g_m</math> (<math>g_m \propto V_{in}</math>)</li> <li>• Bidirectional capability (drain and source are interchangeable)</li> <li>• A near ideal switching device</li> </ul> | <ul style="list-style-type: none"> <li>• High power dissipation</li> <li>• Low input impedance (high drive current)</li> <li>• Low voltage swing logic</li> <li>• Low packing density</li> <li>• Low delay sensitivity to load</li> <li>• High output drive current</li> <li>• High <math>g_m</math> (<math>g_m \propto e^{V_{in}}</math>)</li> <li>• High <math>f_t</math> at low currents</li> <li>• Essentially unidirectional</li> </ul> |





Note: For clarity, the layers have not been drawn transparent but BCCD underlies the entire area and the p<sup>+</sup> base underlies all within its boundary.

**Figure 1-16** Arrangement of BiCMOS npn transistor (Orbit 2 μm CMOS)

are dealt with in Chapter 3. A quick appraisal of Figure 3-13(f) will serve to further illustrate the actual geometry of a BiCMOS bipolar transistor in n-well technology. Since extra design and processing steps are involved, there is an inevitable increase in cost and this is reflected in Figure 1-17, which also includes ECL and GaAs gates for cost comparison.

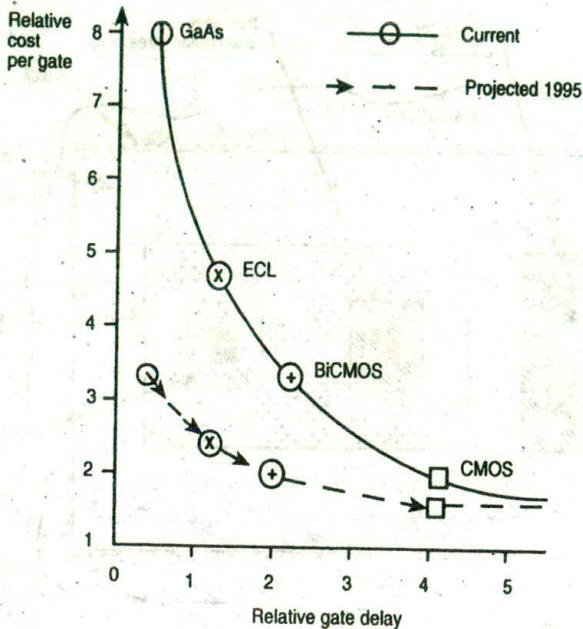


Figure 1-17 Cost versus delay for logic gate

### 1.10.1 BiCMOS fabrication in an n-well process

The basic process steps used are those already outlined for CMOS but with additional process steps and additional masks defining (i) the  $p^+$  base region; (ii)  $n^+$  collector area; and (iii) the buried subcollector (BCCD).

Table 1-3 sets out the process steps for a single poly. single metal CMOS n-well process, showing the additional process steps for the bipolar devices.

### 1.10.2 Some aspects of bipolar and CMOS devices

Clearly there are relative advantages and disadvantages when comparing bipolar technology with CMOS technology. A readily assimilated comparison of some key features was set out in Table 1-2.

It will be seen that there are several advantages if the properties of CMOS and bipolar technologies could be combined. This is achieved to a significant extent in the BiCMOS technology. As in all things, there is a penalty which, in this case, arises from the additional process steps, some loss of packing density and thus higher cost.



**Table 1-3** n-well BiCMOS fabrication process steps

| <i>Single poly. single metal CMOS</i>  | <i>Additional steps for bipolar devices</i>  |
|--|--|
| <ul style="list-style-type: none"> <li>• Form n-well</li> <li>• Delineate active areas</li> <li>• Channel stop</li> <li>• Threshold <math>V_t</math> adjustment</li> <li>• Delineate poly./gate areas</li> <li>• Form <math>n^+</math> active areas</li> <li>• Form <math>p^+</math> active areas</li> <li>• Define contacts</li> <li>• Delineate the metal areas</li> </ul> | <ul style="list-style-type: none"> <li>• Form buried <math>n^+</math> layer (BCCD)</li> <li>• Form deep <math>n^+</math> collector</li> <li>• Form <math>p^+</math> base for bipolars</li> </ul> |

A cost comparison of all current high speed technologies may be assessed from Figure 1-17.

A further advantage which arises from BiCMOS technology is that analog amplifier design is facilitated and improved. High impedance CMOS transistors may be used for the input circuitry while the remaining stages and output drivers are realized using bipolar transistors.

To take maximum advantage of available silicon technologies one might envisage the following mix of technologies in a silicon system:

|        |   |
|--------|---|
| CMOS   | for logic                                   |
| BiCMOS | for I/O and driver circuits                 |
| ECL    | for critical high speed parts of the system |

However, in this text we will not be dealing with the ECL technology.

## 1.11 Production of E-beam masks

All the processes discussed have made use of masks at various stages of fabrication. In many processes, the masks are produced by standard optical techniques and much has been written on the photolithographic processes involved. However, as geometric dimensions shrink, and also to allow for the processing of a number of different chip designs on a single wafer, other techniques are evolving. One popular process employed for this purpose uses an E-beam machine. A rough outline of this type of mask making follows:

1. The starting material consists of chrome-plated glass plates which are coated with an E-beam sensitive resist.
2. The E-beam machine is loaded with the mask description data (MEBES).
3. Plates are loaded into the E-beam machine, where they are exposed with the patterns specified by the customer's mask data.
4. After exposure to the E-beam, the plates are introduced into a developer to bring out the patterns left by the E-beam in the resist coating.
5. The cycle is followed by a bake cycle and a plasma de-summing, which removes the resist residue.
6. The chrome is then etched and the plate is stripped of the remaining E-beam resist.

The advantages of E-beam masks are:

- tighter layer to layer registration;
- smaller feature sizes.

There are two approaches to the design of E-beam machines:

- raster scanning;
- vector scanning.

In the first case, the electron beam scans all possible locations (in a similar fashion to a television display), and a bit map is used to turn the E-beam on and off, depending on whether the particular location being scanned is to be exposed or not.

For vector scanning, the beam is directed only to those locations which are to be exposed. Although this is inherently faster, the data handling involved is more complex.

## 1.12 Observations

This chapter has set the scene by introducing the basically simple MOS transistor structures and the relatively straightforward fabrication processes used in the manufacture of nMOS, CMOS and BiCMOS circuits. We have also attempted to emphasize the revolutionary spread of integrated circuit technology which has, in the short space of 30 years, advanced to a point where we now see highly complex systems completely integrated onto a single chip.

Although this text concentrates on digital circuits and systems, similar techniques can be applied to the design and fabrication of analog devices. Indeed, the trends are toward systems of VLSI (and beyond) complexity which will in future include, on single chips, significant analog interfaces and other appropriate circuitry. This higher level of integration will lead to fewer packages and interconnections and



to more complex systems than today. There will be a marked beneficial effect on cost and reliability of the systems that will be available to all professions and disciplines and in most aspects of everyday life.

Our discussions of fabrication have in some instances simplified the processes used in order to reveal or emphasize the essential features. Indeed, the fabrication of similar devices by different fabricators may vary considerably in detail. This is also the case with the design rules (see Chapter 3) which are specified by the fabricator. Design rules will be introduced via the concept of 'lambda-based' rules, which are a result of the work of Mead and Conway, and although not producing the tightest layouts, these rules are acceptable to many fabricators. A study of lambda-based rules also provides a good way of absorbing the essential concepts underlying any set of design rules. However, the text also gives an up-to-date set of real world 'micron-based' rules for  $2\ \mu\text{m}$  and for  $1.2\ \mu\text{m}$  n-well CMOS technologies which may be used when the designer reaches an acceptable level of competence. The  $2\ \mu\text{m}$  rule set is for a BiCMOS process and thus also provides for bipolar npn transistors. It must be noted here that '2  $\mu\text{m}$  technology', for example, means that the minimum line width (and, consequently, the typical feature size of the geometry) of the chip layout will be  $2\ \mu\text{m}$ .

In order to understand the basic features MOS and BiCMOS IC technologies, we must now look into the basic electrical properties.

# 2

## Basic electrical properties of MOS and BiCMOS circuits

*There is no virtue in not knowing what can be known.*

Aldous Huxley

### Objectives

If design is to be effectively carried out, or indeed if the performance of circuits realized in MOS technology is to be properly understood, then the practitioner must have a sound understanding of the MOS active devices.

This chapter establishes the basic characteristics of MOS transistor and examines various possibilities for configuring inverter circuits. In the case of nMOS circuits the need for and values of the ratio rules are established.

Discussion then extends to the characteristics of BiCMOS transistors and the ensuing inverter circuitry.

Finally, aspects of latch-up are considered for CMOS and BiCMOS devices. Having introduced the MOS transistor and the processes used to produce it, we are now in a position to gain some understanding of the electrical characteristics of the basic MOS circuits — enhancement and depletion mode transistors and inverters. Our considerations will be based on reasonable approximations so that the essential features can be evaluated and illustrated in a concise and easily absorbed manner. VLSI designers should have a good knowledge of the behavior of the circuits they are designing or designing with. Even if large systems are being designed, using computer-aided design processes, it is essential that the designs be based on a sound foundation of understanding if those systems are to meet performance specifications.



The following expressions and discussion relate directly to nMOS transistors, but pMOS expressions are also given where appropriate and, generally, a reversal of voltage and current polarities of nMOS expressions and the exchange of  $\mu_n$  for  $\mu_p$  and electrons for holes will yield pMOS from nMOS expressions.

We will then briefly discuss some bipolar transistor characteristics which are relevant to an understanding of BiCMOS circuits. Bipolar transistor parameters are also compared with comparable parameters for CMOS transistors.

## 2.1 Drain-to-source current $I_{ds}$ versus voltage $V_{ds}$ relationships

The whole concept of the MOS transistor evolves from the use of a voltage on the gate to induce a charge in the channel between source and drain, which may then be caused to move from source to drain under the influence of an electric field created by voltage  $V_{ds}$  applied between drain and source. Since the charge induced is dependent on the gate to source voltage  $V_{gs}$ , then  $I_{ds}$  is dependent on both  $V_{gs}$  and  $V_{ds}$ . Consider a structure, as in Figure 2-1, in which electrons will flow source to drain:

$$I_{ds} = -I_{sd} = \frac{\text{Charge induced in channel } (Q_C)}{\text{Electron transit time } (\tau)} \quad (2.1)$$

First, transit time:

$$\tau_{sd} = \frac{\text{Length of channel } (L)}{\text{Velocity } (v)}$$

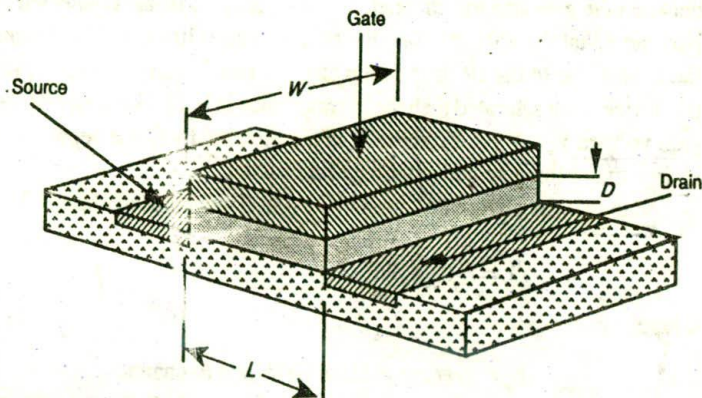


Figure 2-1 nMOS transistor structure

but velocity

$$v = \mu E_{ds}$$

where

$$\begin{aligned} \mu &= \text{electron or hole mobility (surface)} \\ E_{ds} &= \text{electric field (drain to source)} \end{aligned}$$

Now

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

so that

$$v = \frac{\mu V_{ds}}{L}$$

Thus

$$\tau_{sd} = \frac{L^2}{\mu V_{ds}} \quad (2.2)$$

Typical values of  $\mu$  at room temperature are:

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

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

### 2.1.1 The non-saturated region

Charge induced in channel due to gate voltage is due to the voltage difference between the gate and the channel,  $V_{gs}$  (assuming substrate connected to source). Now note that the voltage along the channel varies linearly with distance  $X$  from the source due to the IR drop in the channel (see Figure 1-5) and assuming that the device is not saturated then the average value is  $V_{ds}/2$ . Furthermore, the effective gate voltage  $V_g = V_{gs} - V_t$  where  $V_t$  is the threshold voltage needed to invert the charge under the gate and establish the channel.

Note that the charge/unit area =  $E_g \epsilon_{ins} \epsilon_0$ . Thus induced charge

$$Q_c = E_g \epsilon_{ins} \epsilon_0 WL$$

where

$$\begin{aligned} E_g &= \text{average electric field gate to channel} \\ \epsilon_{ins} &= \text{relative permittivity of insulation between gate and channel} \\ \epsilon_0 &= \text{permittivity of free space} \end{aligned}$$

(Note:  $\epsilon_0 = 8.85 \times 10^{-14} \text{ F cm}^{-1}$ ;  $\epsilon_{ins} \doteq 4.0$  for silicon dioxide)



Now

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

where  $D$  = oxide thickness.

Thus

$$Q_c = \frac{WL\epsilon_{ins}\epsilon_0}{D} \left( (V_{gs} - V_t) - \frac{V_{ds}}{2} \right) \quad (2.3)$$

Now, combining equations (2.2) and (2.3) in equation (2.1), we have

$$I_{ds} = \frac{\epsilon_{ins}\epsilon_0\mu}{D} \frac{W}{L} \left( (V_{gs} - V_t) - \frac{V_{ds}}{2} \right) V_{ds}$$

or

$$I_{ds} = K \frac{W}{L} \left( (V_{gs} - V_t) V_{ds} - \frac{V_{ds}^2}{2} \right) \quad (2.4)$$

in the non-saturated or resistive region where  $V_{ds} < V_{gs} - V_t$  and

$$K = \frac{\epsilon_{ins}\epsilon_0\mu}{D}$$

The factor  $W/L$  is, of course, contributed by the geometry and it is common practice to write

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

so that

$$I_{ds} = \beta \left( (V_{gs} - V_t) V_{ds} - \frac{V_{ds}^2}{2} \right) \quad (2.4a)$$

which is an alternative form of equation 2.4.

Noting that gate/channel capacitance

$$C_g = \frac{\epsilon_{ins}\epsilon_0 WL}{D} \quad (\text{parallel plate})$$

we also have

$$K = \frac{C_g\mu}{WL}$$

so that

$$I_{ds} = \frac{C_g \mu}{L^2} \left( (V_{gs} - V_t) V_{ds} - \frac{V_{ds}^2}{2} \right) \quad (2.4b)$$

which is a further alternative form of equation 2.4.

Sometimes it is convenient to use *gate capacitance per unit area*  $C_0$  (which is often denoted  $C_{ox}$ ) rather than  $C_g$  in this and other expressions. Noting that

$$C_g = C_0 WL$$

we may also write

$$I_{ds} = C_0 \mu \frac{W}{L} \left( (V_{gs} - V_t) V_{ds} - \frac{V_{ds}^2}{2} \right) \quad (2.4c)$$

### 2.1.2 The saturated region

*Saturation* begins when  $V_{ds} = V_{gs} - V_t$  since at this point the IR drop in the channel equals the effective gate to channel voltage at the drain and we may assume that the current remains fairly constant as  $V_{ds}$  increases further. Thus

$$I_{ds} = K \frac{W}{L} \frac{(V_{gs} - V_t)^2}{2} \quad (2.5)$$

or, we may write

$$I_{ds} = \frac{\beta}{2} (V_{gs} - V_t)^2 \quad (2.5a)$$

or

$$I_{ds} = \frac{C_g \mu}{2L^2} (V_{gs} - V_t)^2 \quad (2.5b)$$

We may also write

$$I_{ds} = C_0 \mu \frac{W}{2L} (V_{gs} - V_t)^2 \quad (2.5c)$$

The expressions derived for  $I_{ds}$  hold for both enhancement and depletion mode devices, but it should be noted that the threshold voltage for the nMOS depletion mode device (denoted as  $V_{td}$ ) is *negative*.

Typical characteristics for nMOS transistors are given in Figure 2-2. pMOS transistor characteristics are similar, with suitable reversal of polarity.



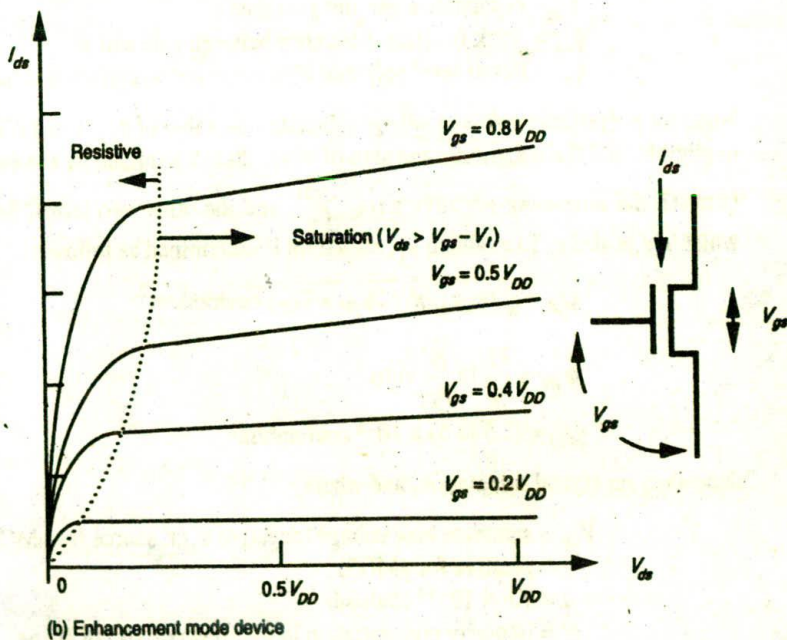
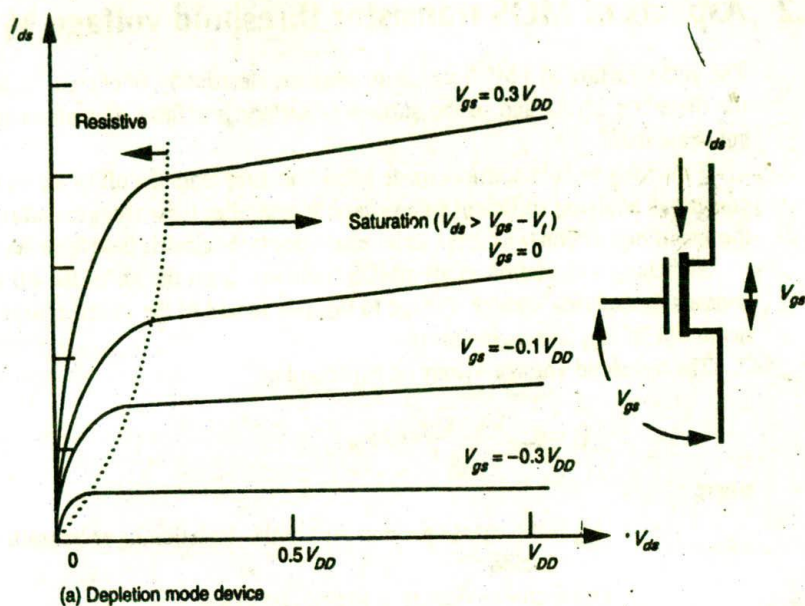


Figure 2-2 MOS transistor characteristics

## 2.2 Aspects of MOS transistor threshold voltage $V_t$

The gate structure of a MOS transistor consists, electrically, of charges stored in the dielectric layers and in the surface to surface interfaces as well as in the substrate itself.

Switching an enhancement mode MOS transistor from the off to the on state consists in applying sufficient gate voltage to neutralize these charges and enable the underlying silicon to undergo an inversion due to the electric field from the gate.

Switching a depletion mode nMOS transistor from the on to the off state consists in applying enough voltage to the gate to add to the stored charge and invert the 'n' implant region to 'p'.

The threshold voltage  $V_t$  may be expressed as:

$$V_t = \phi_{ms} \frac{Q_B - Q_{SS}}{C_0} + 2\phi_{fN} \quad (2.6)$$

where

$Q_B$  = the charge per unit area in the depletion layer beneath the oxide

$Q_{SS}$  = charge density at Si:SiO<sub>2</sub> interface

$C_0$  = capacitance per unit gate area

$\phi_{ms}$  = work function difference between gate and Si

$\phi_{fN}$  = Fermi level potential between inverted surface and bulk Si.

Now, for polysilicon gate and silicon substrate, the value of  $\phi_{ms}$  is negative but negligible, and the magnitude and sign of  $V_t$  are thus determined by the balance between the remaining negative term  $\frac{-Q_{SS}}{C_0}$  and the other two terms, both of which are positive. To evaluate  $V_t$ , each term is determined as follows:

$$Q_B = \sqrt{2\epsilon_0\epsilon_{Si}qN(2\phi_{fN} + V_{SB})} \text{ coulomb/m}^2$$

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

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

depending on crystal orientation, and where

$V_{SB}$  = substrate bias voltage (negative w.r.t. source for nMOS, positive for pMOS)

$q$  =  $1.6 \times 10^{-19}$  coulomb

$N$  = impurity concentration in the substrate ( $N_A$  or  $N_D$  as appropriate)

$\epsilon_{Si}$  = relative permittivity of silicon  $\doteq 11.7$

$n_i$  = intrinsic electron concentration ( $1.6 \times 10^{10}/\text{cm}^3$  at 300°K)

$k$  = Boltzmann's constant =  $1.4 \times 10^{-23}$  joule/°K



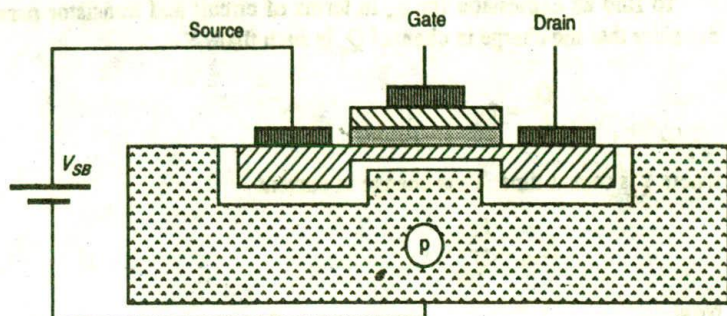


Figure 2-3 Body effect (nMOS device shown)

The *body effects* may also be taken into account since the substrate may be biased with respect to the source, as shown in Figure 2-3.

Increasing  $V_{SB}$  causes the channel to be depleted of charge carriers and thus the threshold voltage is raised.

Change in  $V_t$  is given by  $\Delta V_t \doteq \gamma(V_{SB})^{1/2}$  where  $\gamma$  is a constant which depends on substrate doping so that the more lightly doped the substrate, the smaller will be the body effect.

Alternatively, we may write

$$V_t = V_t(0) + \left( \frac{D}{\epsilon_{ins} \epsilon_0} \right) \sqrt{2\epsilon_0 \epsilon_{Si} QN} \cdot (V_{SB})^{1/2}$$

where  $V_t(0)$  is the threshold voltage for  $V_{SB} = 0$ .

To establish the magnitude of such effects, typical figures for  $V_t$  are as follows:

For nMOS enhancement mode transistors:

$$\left. \begin{array}{l} V_{SB} = 0 \text{ V}; V_t = 0.2V_{DD} (= +1 \text{ V for } V_{DD} = +5 \text{ V}) \\ V_{SB} = 5 \text{ V}; V_t = 0.3V_{DD} (= +1.5 \text{ V for } V_{DD} = +5 \text{ V}) \end{array} \right\} \begin{array}{l} \text{Similar but} \\ \text{negative values} \\ \text{for pMOS} \end{array}$$

For nMOS depletion mode transistors:

$$\left. \begin{array}{l} V_{SB} = 0 \text{ V}; V_{td} = -0.7V_{DD} (= -3.5 \text{ V for } V_{DD} = +5 \text{ V}) \\ V_{SB} = 5 \text{ V}; V_{td} = -0.6V_{DD} (= -3.0 \text{ V for } V_{DD} = +5 \text{ V}) \end{array} \right\}$$

## 2.3 MOS transistor transconductance $g_m$ and output conductance $g_{ds}$

Transconductance expresses the relationship between output current  $I_{ds}$  and the input voltage  $V_{gs}$  and is defined as

$$g_m = \frac{\delta I_{ds}}{\delta V_{gs}} \Big|_{V_{ds}} = \text{constant}$$

To find an expression for  $g_m$  in terms of circuit and transistor parameters, consider that the charge in channel  $Q_c$  is such that

$$\frac{Q_c}{I_{ds}} = \tau$$

where  $\tau$  is transit time. Thus change in current

$$\delta I_{ds} = \frac{\delta Q_c}{\tau_{ds}}$$

Now

$$\tau_{ds} = \frac{L^2}{\mu V_{ds}} \quad (\text{from 2.2})$$

Thus

$$\delta I_{ds} = \frac{\delta Q_c V_{ds} \mu}{L^2}$$

but change in charge

$$\delta Q_c = C_g \delta V_{gs}$$

so that

$$\delta I_{ds} = \frac{C_g \delta V_{gs} \mu V_{ds}}{L^2}$$

Now

$$g_m = \frac{\delta I_{ds}}{\delta V_{gs}} = \frac{C_g \mu V_{ds}}{L^2}$$

In saturation

$$V_{ds} = V_{gs} - V_t$$

$$g_m = \frac{C_g \mu}{L^2} (V_{gs} - V_t) \quad (2.7)$$

and substituting for  $C_g = \frac{\epsilon_{ins} \epsilon_0 WL}{D}$

$$g_m = \frac{\mu \epsilon_{ins} \epsilon_0 W}{D L} (V_{gs} - V_t) \quad (2.7a)$$

Alternatively,

$$g_m = \beta (V_{gs} - V_t)$$



It is possible to increase the  $g_m$  of a MOS device by increasing its width. However, this will also increase the input capacitance and area occupied.

A reduction in the channel length results in an increase in  $\omega_0$  owing to the higher  $g_m$ . However, the gain of the MOS device decreases owing to the strong degradation of the output resistance  $= 1/g_{ds}$ .

The output conductance  $g_{ds}$  can be expressed by

$$g_{ds} = \frac{\delta I_{ds}}{\delta V_{gs}} = \lambda \cdot I_{ds} \alpha \left( \frac{1}{L} \right)^2$$

Here the strong dependence on the channel length is demonstrated as

$$\lambda \alpha \left( \frac{1}{L} \right) \text{ and } I_{ds} \alpha \left( \frac{1}{L} \right)$$

for the MOS device.

## 2.4 MOS transistor figure of merit $\omega_0$

An indication of frequency response may be obtained from the parameter  $\omega_0$  where

$$\omega_0 = \frac{g_m}{C_g} = \frac{\mu}{L^2} (V_{gs} - V_t) \left( = \frac{1}{\tau_{sd}} \right) \quad (2.8)$$

This shows that switching speed depends on gate voltage above threshold and on carrier mobility and inversely as the square of channel length. A fast circuit requires that  $g_m$  be as high as possible.

Electron mobility on a (100) oriented n-type inversion layer surface ( $\mu_n$ ) is larger than that on a (111) oriented surface, and is in fact about three times as large as hole mobility on a (111) oriented p-type inversion layer. Surface mobility is also dependent on the effective gate voltage ( $V_{gs} - V_t$ ).

For faster nMOS circuits, then, one would choose a (100) oriented p-type substrate in which the inversion layer will have a surface carrier mobility  $\mu_n \doteq 650 \text{ cm}^2/\text{V sec}$  at room temperature.

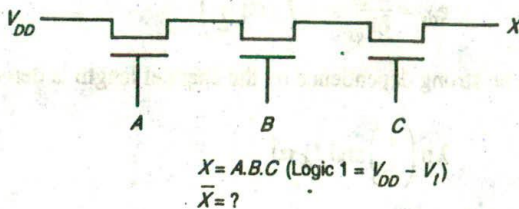
Compare this with the typical bulk mobilities

$$\begin{aligned} \mu_n &= 1250 \text{ cm}^2/\text{V sec} \\ \mu_p &= 480 \text{ cm}^2/\text{V sec} \end{aligned}$$

from which it will be seen that  $\frac{\mu_s}{\mu} \doteq 0.5$  (where  $\mu_s$  = surface mobility and  $\mu$  = bulk mobility).

## 2.5 The pass transistor

Unlike bipolar transistors, the isolated nature of the gate allows MOS transistors to be used as switches in series with lines carrying logic levels in a way that is similar to the use of relay contacts. This application of the MOS device is called the *pass transistor* and switching logic arrays can be formed — for example, an *And* array as in Figure 2-4.



Note: Means must exist so that  $X$  assumes ground potential when  $A + B + C = 0$ .

Figure 2-4 Pass transistor *And* gate

## 2.6 The nMOS inverter

A basic requirement for producing a complete range of logic circuits is the inverter. This is needed for restoring logic levels, for *Nand* and *Nor* gates, and for sequential and memory circuits of various forms. In the treatment of the inverter used in this section, the authors wish to acknowledge the influence of material previously published by Mead and Conway.

The basic inverter circuit requires a transistor with source connected to ground and a load resistor of some sort connected from the drain to the positive supply rail  $V_{DD}$ . The output is taken from the drain and the input applied between gate and ground.

Resistors are not conveniently produced on the silicon substrate; even modest values occupy excessively large areas so that some other form of load resistance is required. A convenient way to solve this problem is to use a depletion mode transistor as the load, as shown in Figure 2-5.

Now:

- With no current drawn from the output, the currents  $I_{ds}$  for both transistors must be equal.
- For the depletion mode transistor, the gate is connected to the source so it is always on and only the characteristic curve  $V_{gs} = 0$  is relevant.
- In this configuration the depletion mode device is called the pull-up (p.u.) and the enhancement mode device the pull-down (p.d.) transistor.



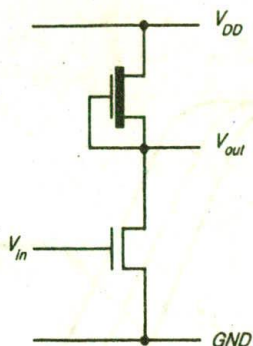
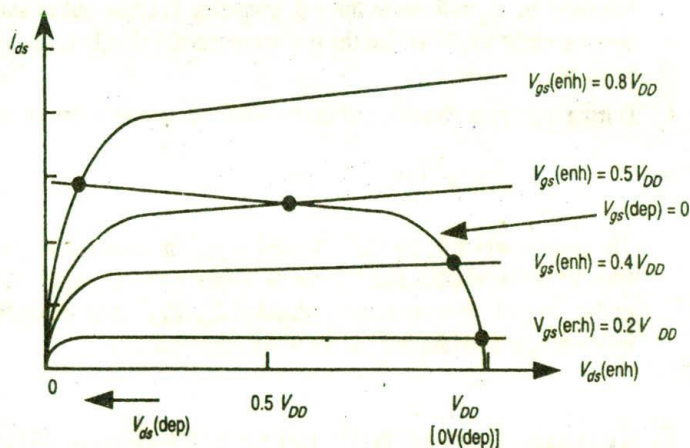


Figure 2-5 nMOS inverter

- To obtain the inverter transfer characteristic we superimpose the  $V_{gs} = 0$  depletion mode characteristic curve on the family of curves for the enhancement mode device, noting that maximum voltage across the enhancement mode device corresponds to minimum voltage across the depletion mode transistor.
- The points of intersection of the curves as in Figure 2-6 give points on the transfer characteristic, which is of the form shown in Figure 2-7.



$$V_{ds}(\text{enh}) = V_{DD} - V_{ds}(\text{dep}) = V_{out}$$

$V_{gs}(\text{enh}) = V_{in}$  ... intersection points give transfer characteristic

Figure 2-6 Derivation of nMOS inverter transfer characteristic

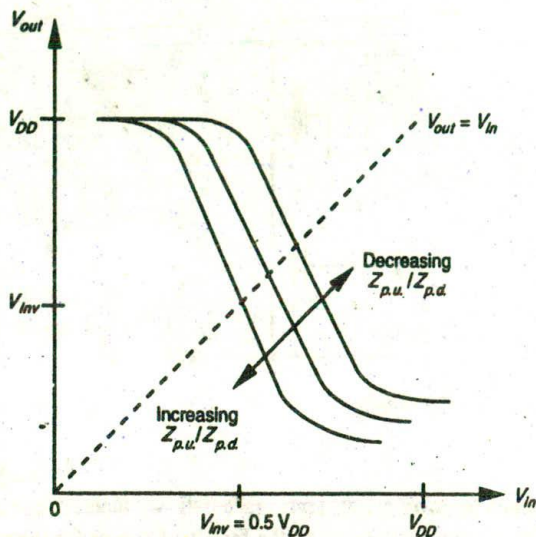


Figure 2-7 nMOS inverter transfer characteristic

- Note that as  $V_{in}$  ( $= V_{gs}$  p.d. transistor) exceeds the p.d. threshold voltage current begins to flow. The output voltage  $V_{out}$  thus decreases and the subsequent increases in  $V_{in}$  will cause the p.d. transistor to come out of saturation and become resistive. Note that the p.u. transistor is initially resistive as the p.d. turns on.
- During transition, the slope of the transfer characteristic determines the gain:

$$\text{Gain} = \frac{\delta V_{out}}{\delta V_{in}}$$

- The point at which  $V_{out} = V_{in}$  is denoted as  $V_{inv}$  and it will be noted that the transfer characteristics and  $V_{inv}$  can be shifted by variation of the ratio of pull-up to pull-down resistances (denoted  $Z_{p.u.}/Z_{p.d.}$  where  $Z$  is determined by the length to width ratio of the transistor in question).

## 2.7 Determination of pull-up to pull-down ratio ( $Z_{p.u.}/Z_{p.d.}$ ) for an nMOS inverter driven by another nMOS inverter

Consider the arrangement in Figure 2-8 in which an inverter is driven from the output of another similar inverter. Consider the depletion mode transistor for



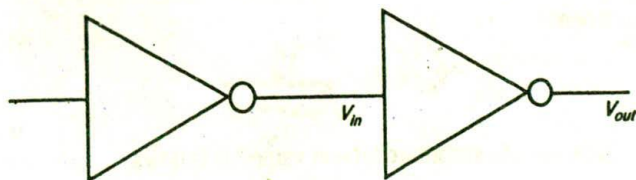


Figure 2-8 nMOS inverter driven directly by another inverter

which  $V_{gs} = 0$  under all conditions, and further assume that in order to cascade inverters without degradation of levels we are aiming to meet the requirement

$$V_{in} = V_{out} = V_{inv}$$

For equal margins around the inverter threshold, we set  $V_{inv} = 0.5V_{DD}$ . At this point both transistors are in saturation and

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

In the depletion mode

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

and in the enhancement mode

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

Equating (since currents are the same) we have

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

where  $W_{p.d.}$ ,  $L_{p.d.}$ ,  $W_{p.u.}$ , and  $L_{p.u.}$  are the widths and lengths of the pull-down and pull-up transistors respectively.

Now write

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

we have

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

whence

$$V_{inv} = V_t - \frac{V_{td}}{\sqrt{Z_{p,u.} / Z_{p,d.}}} \quad (2.9)$$

Now we can substitute typical values as follows

$$V_t = 0.2V_{DD}; V_{td} = -0.6V_{DD}$$

$$V_{inv} = 0.5V_{DD} \text{ (for equal margins)}$$

thus, from equation (2.9)

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

whence

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

and thus

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

for an inverter directly driven by an inverter.

## 2.8 Pull-up to pull-down ratio for an nMOS inverter driven through one or more pass transistors

Now consider the arrangement of Figure 2-9 in which the input to inverter 2 comes from the output of inverter 1 but passes through one or more nMOS transistors used as switches in series (called *pass transistors*).

We are concerned that connection of pass transistors in series will degrade the logic 1 level into inverter 2 so that the output will not be a proper logic 0 level. The critical condition is when point A is at 0 volts and B is thus at  $V_{DD}$ , but the voltage into inverter 2 at point C is now reduced from  $V_{DD}$  by the threshold

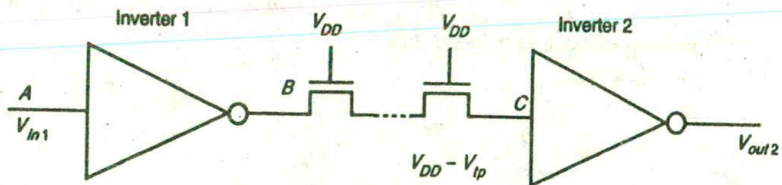


Figure 2-9 Pull-up to pull-down ratios for inverting logic coupled by pass transistors



voltage of the series pass transistor. With all pass transistor gates connected to  $V_{DD}$  (as shown in Figure 2-8), there is a loss of  $V_{tp}$ , however many are connected in series, since no static current flows through them and there can be no voltage drop in the channels. Therefore, the input voltage to inverter 2 is

$$V_{in2} = V_{DD} - V_{tp}$$

where

$$V_{tp} = \text{threshold voltage for a pass transistor}$$

We must now ensure that for this input voltage we get out the same voltage as would be the case for inverter 1 driven with input =  $V_{DD}$ .

Consider inverter 1 (Figure 2-10(a)) with input =  $V_{DD}$ . If the input is at  $V_{DD}$ , then the p.d. transistor  $T_2$  is conducting but with a low voltage across it; therefore, it is in its resistive region represented by  $R_1$  in Figure 2-10. Meanwhile, the p.u. transistor  $T_1$  is in saturation and is represented as a current source.

For the p.d. transistor

$$I_{ds} = K \frac{W_{p.d.1}}{L_{p.d.1}} \left( (V_{DD} - V_t) V_{ds1} - \frac{V_{ds1}^2}{2} \right) \quad (\text{from 2.4})$$

Therefore

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

Note that  $V_{ds1}$  is small and  $\frac{V_{ds1}}{2}$  may be ignored.

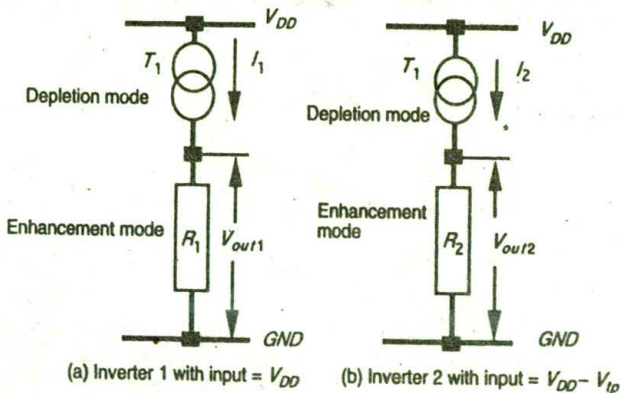


Figure 2-10 Equivalent circuits of inverters 1 and 2

Thus

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

Now, for depletion mode p.u. in saturation with  $V_{gs} = 0$

$$I_1 = I_{ds} = K \frac{W_{p.u.1} (-V_{td})^2}{L_{p.u.1} 2} \quad (\text{from 2.5})$$

The product

$$I_1 R_1 = V_{out1}$$

Thus

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

Consider inverter 2 (Figure 2-10(b)) when input =  $V_{DD} - V_{tp}$ . As for inverter 1

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

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

whence

$$V_{out2} = I_2 R_2 = \frac{Z_{p.d.2}}{Z_{p.u.2}} \left( \frac{1}{(V_{DD} - V_{tp} - V_t)} \right) \frac{(-V_{td})^2}{2}$$

If inverter 2 is to have the same output voltage under these conditions then  $V_{out1} = V_{out2}$ . That is

$$I_1 R_1 = I_2 R_2$$

Therefore

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

Taking typical values

$$V_t = 0.2 V_{DD}$$

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

$$\frac{Z_{p.u.2}}{Z_{p.d.2}} = \frac{Z_{p.u.1}}{Z_{p.d.1}} \frac{0.8}{0.5}$$



Therefore

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

Summarizing for an nMOS inverter:

- An inverter driven directly from the output of another should have a  $Z_{p.u.}/Z_{p.d.}$  ratio of  $\geq 4/1$ .
- An inverter driven through one or more pass transistors should have a  $Z_{p.u.}/Z_{p.d.}$  ratio of  $\geq 8/1$ .

*Note:* It is the driven, *not* the driver, whose ratio is affected.

## 2.9 Alternative forms of pull-up

Up to now we have assumed that the inverter circuit has a depletion mode pull-up transistor as its load. There are, however, at least four possible arrangements:

1. *Load resistance  $R_L$*  (Figure 2-11). This arrangement is not often used because of the large space requirements of resistors produced in a silicon substrate.
2. *nMOS depletion mode transistor pull-up* (Figure 2-12).
  - (a) Dissipation is high since rail to rail current flows when  $V_{in} = \text{logical } 1$ .
  - (b) Switching of output from 1 to 0 begins when  $V_{in}$  exceeds  $V_t$  of p.d. device.
  - (c) When switching the output from 1 to 0, the p.u. device is non-saturated initially and this presents lower resistance through which to charge capacitive loads.

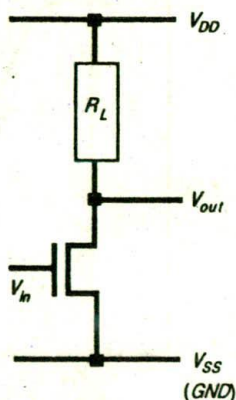


Figure 2-11 Resistor pull-up

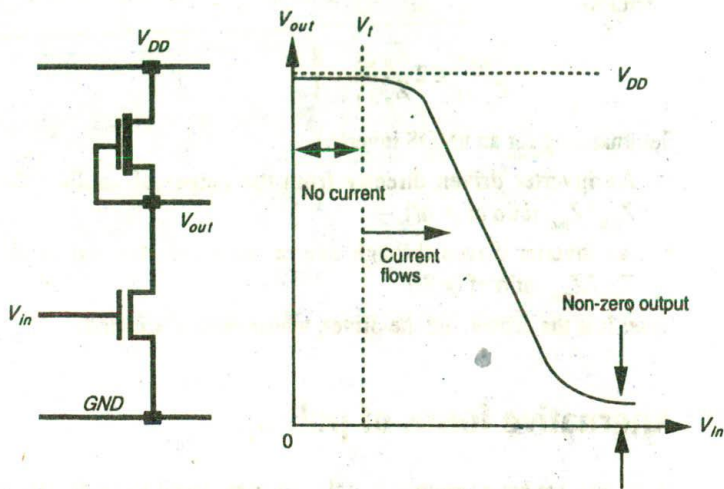


Figure 2-12 nMOS depletion mode transistor pull-up and transfer characteristic

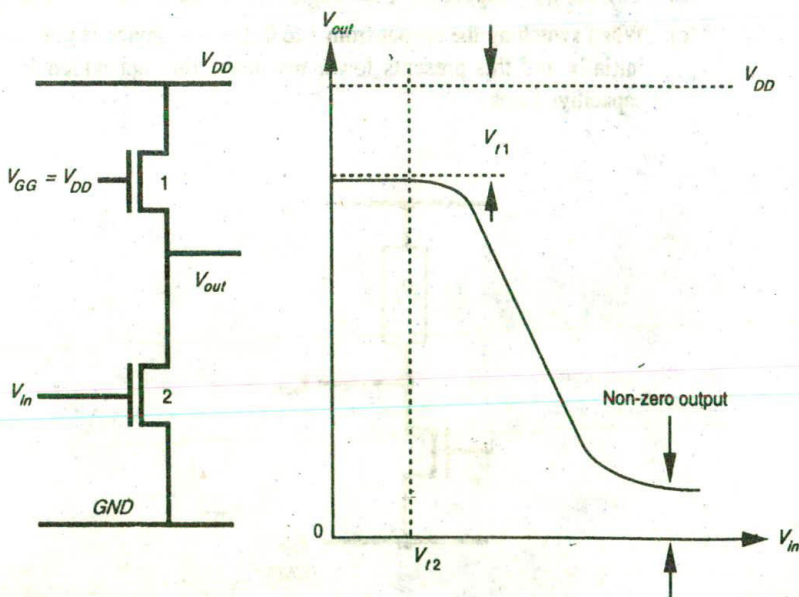


Figure 2-13 nMOS enhancement mode pull-up and transfer characteristic



3. *nMOS enhancement mode pull-up* (Figure 2-13).
  - (a) Dissipation is high since current flows when  $V_{in} = \text{logical 1}$  ( $V_{GG}$  is returned to  $V_{DD}$ ).
  - (b)  $V_{out}$  can never reach  $V_{DD}$  (logical 1) if  $V_{GG} = V_{DD}$  as is normally the case.
  - (c)  $V_{GG}$  may be derived from a switching source, for example, one phase of a clock, so that dissipation can be greatly reduced.
  - (d) If  $V_{GG}$  is higher than  $V_{DD}$  then an extra supply rail is required.
4. *Complementary transistor pull-up* (CMOS) (Figure 2-14).
  - (a) No current flow either for logical 0 or for logical 1 inputs.
  - (b) Full logical 1 and 0 levels are presented at the output.
  - (c) For devices of similar dimensions the p-channel is slower than the n-channel device.

## 2.10 The CMOS inverter

The general arrangement and characteristics are illustrated in Figure 2-14. We have seen (equations 2.4 and 2.5) that the current/voltage relationships for the MOS transistor may be written

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

in the resistive region, or

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

in saturation. In both cases the factor  $K$  is a technology-dependent parameter such that

$$K = \frac{\epsilon_{ins} \epsilon_0 \mu}{D}$$

The factor  $W/L$  is, of course, contributed by the geometry and it is common practice to write

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

so that, for example

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

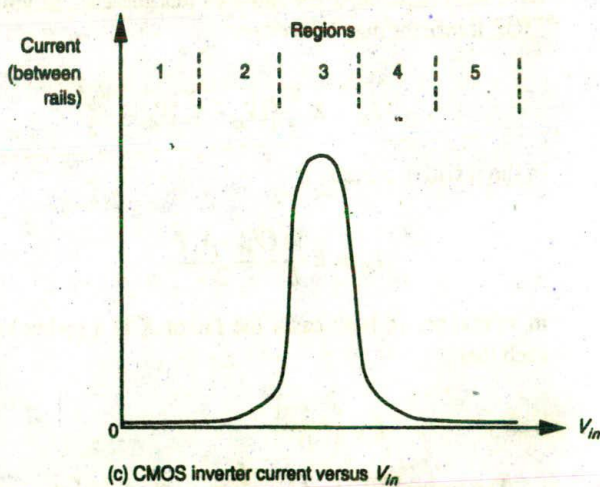
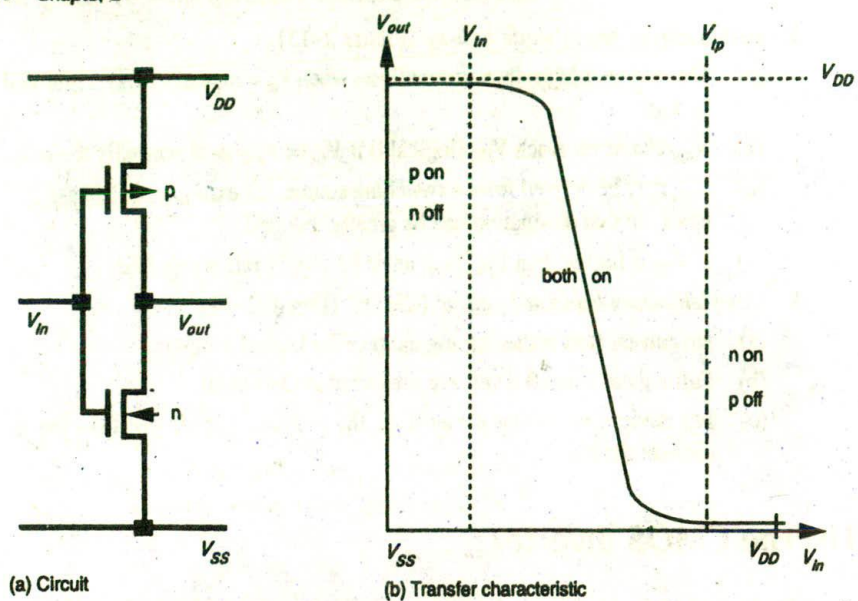


Figure 2-14 Complementary transistor pull-up (CMOS)

in saturation, and where  $\beta$  may be applied to both nMOS and pMOS transistors as follows

$$\beta_n = \frac{\epsilon_{ins} \epsilon_0 \mu_n}{D} \frac{W_n}{L_n}$$

$$\beta_p = \frac{\epsilon_{ins} \epsilon_0 \mu_p}{D} \frac{W_p}{L_p}$$

where  $W_n$  and  $L_n$ ,  $W_p$  and  $L_p$  are the n- and p-transistor dimensions respectively. With regard to Figures 2-14(b) and 2-14(c), it may be seen that the CMOS inverter has five distinct regions of operation.

Considering the static conditions first, it may be seen that in *region 1* for which  $V_{in} = \text{logic } 0$ , we have the p-transistor fully turned on while the n-transistor is fully turned off. Thus no current flows through the inverter and the output is directly connected to  $V_{DD}$  through the p-transistor. A good logic 1 output voltage is thus present at the output.

In *region 5*  $V_{in} = \text{logic } 1$ , the n-transistor is fully on while the p-transistor is fully off. Again, no current flows and a good logic 0 appears at the output.

In *region 2* the input voltage has increased to a level which just exceeds the threshold voltage of the n-transistor. The n-transistor conducts and has a large voltage between source and drain; so it is in saturation. The p-transistor is also conducting but with only a small voltage across it, it operates in the unsaturated resistive region. A small current now flows through the inverter from  $V_{DD}$  to  $V_{SS}$ . If we wish to analyze the behavior in this region, we equate the p-device resistive region current with the n-device saturation current and thus obtain the voltage and current relationships.

*Region 4* is similar to *region 2* but with the roles of the p- and n-transistors reversed. However, the current magnitudes in *regions 2* and *4* are small and most of the energy consumed in switching from one state to the other is due to the larger current which flows in *region 3*.

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

The currents (with regard to Figure 2-14(c) in each device must be the same since the transistors are in series, so we may write

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

where

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

and

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

from whence we can express  $V_{in}$  in terms of the  $\beta$  ratio and the other circuit voltages and currents

$$V_{in} = \frac{V_{DD} + V_p + V_m (\beta_n / \beta_p)^{1/2}}{1 + (\beta_n / \beta_p)^{1/2}} \quad (2.10)$$

Since both transistors are in saturation, they act as current sources so that the equivalent circuit in this region is two current sources in series between  $V_{DD}$  and



$V_{SS}$  with the output voltage coming from their common point. The region is inherently unstable in consequence and the changeover from one logic level to the other is rapid.

If  $\beta_n = \beta_p$  and if  $V_{in} = -V_{ip}$ , then from equation 2.10

$$V_{in} = 0.5 V_{DD}$$

This implies that the changeover between logic levels is symmetrically disposed about the point at which

$$V_{in} = V_{out} = 0.5 V_{DD}$$

since only at this point will the two  $\beta$  factors be equal. But for  $\beta_n = \beta_p$  the device geometries must be such that

$$\mu_p W_p / L_p = \mu_n W_n / L_n$$

Now the mobilities are inherently unequal and thus it is necessary for the width to length ratio of the p-device to be two to three times that of the n-device, namely

$$W_p / L_p \doteq 2.5 W_n / L_n$$

However, it must be recognized that mobility  $\mu$  is affected by the transverse electric field in the channel and is thus dependent on  $V_{gs}$  (and thus on  $V_{in}$  in this case). It has been shown empirically that the actual mobility is

$$\mu = \mu_z (1 - \phi(V_{gs} - V_t))^{-1}$$

$\phi$  is a constant approximately equal to 0.05,  $V_t$  includes any body effect, and  $\mu_z$  is the mobility with zero transverse field. Thus a  $\beta$  ratio of 1 will only hold good around the point of symmetry when  $V_{out} = V_{in} = 0.5 V_{DD}$ .

The  $\beta$  ratio is often unimportant in many configurations and in most cases minimum size transistor geometries are used for both n- and p-devices. Figure 2-15 indicates the trends in the transfer characteristic as the ratio is varied. The changes indicated in the figure would be for quite large variations in  $\beta$  ratio (e.g. up to 10:1) and the ratio is thus not too critical in this respect.

## 2.11 MOS transistor circuit model

The MOS transistor can be modeled with varying degrees of complexity. However, a consideration of the actual physical construction of the device (as in Figure 2-16) leads to some understanding of the various components of the model.

Notes:  $C_{GC}$  = gate to channel capacitance  
 $C_{GS}$  = gate to source capacitance  
 $C_{GD}$  = gate to drain capacitance } Small for self-aligning nMOS process

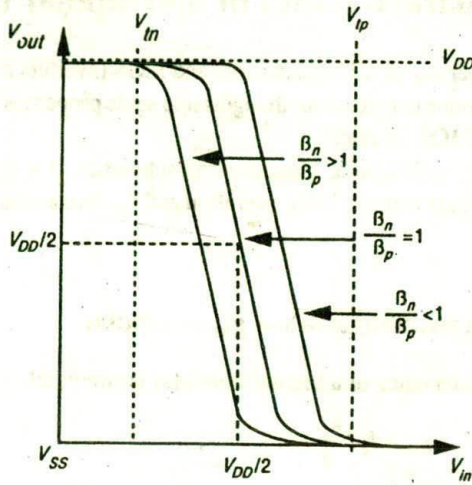


Figure 2-15 Trends in transfer characteristic with  $\beta$  ratio

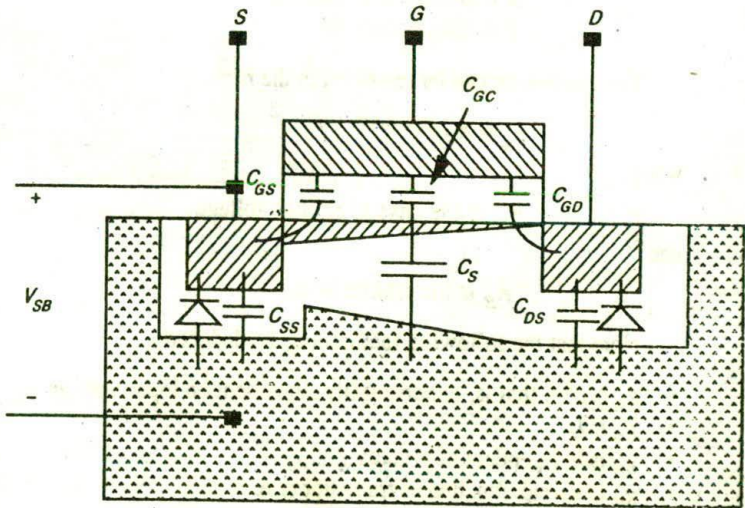


Figure 2-16 nMOS transistor model

Remaining capacitances are associated with the depletion layer and are voltage dependent. Note that  $C_{SS}$  indicates source-to-substrate,  $C_{DS}$  drain-to-substrate and  $C_S$  channel-to-substrate capacitances.

A-03158

## 2.12 Some characteristics of npn bipolar transistors

The key properties of MOS transistors and MOS inverters having been covered, it is now desirable to extend our thoughts into some properties of bipolar transistors and into BiCMOS inverters.

In dealing with bipolar transistor characteristics, it will be assumed that the reader is familiar with the basic operation and the fundamental aspects of bipolar transistors.

### 2.12.1 Transconductance $g_m$ — bipolar

The transconductance of a bipolar transistor is commonly presented as

$$g_m = I_c / \frac{kT}{q}$$

where

$I_c$  = collector current

$q$  = electron charge

$k$  = Boltzmann's constant

$T$  = temperature °K

The expression can be rewritten in the form

$$g_m \propto A_E e^{V_{be}/(q/kT)}$$

where

$V_{be}$  is the base to emitter voltage

and

$A_E$  is the emitter area.

Note that the following factors may be deduced

- $g_m \propto e^{V_{be}}$ , that is, exponentially dependent on input voltage  $V_{be}$
- $g_m \propto I_c$
- $g_m$  is independent of process
- $g_m$  is a weak function of transistor size.

Remembering that, for MOS transistors

$$g_m = \frac{\mu \epsilon_{ins} \epsilon_0 W}{D L} (V_{gs} - V_t)$$

where

$D$  = oxide thickness (often denoted  $t_{ox}$ )

Comparisons can be made between MOS and bipolar transistor  $g_m$  as follows:



1. For  $I_c = I_{ds}$  the difference between the thermal voltage ( $kT/q$ ) and the effective gate voltage ( $V_{gs} - V_t$ ) introduces a large difference in transconductance.
2. If inputs are controlled by equal amounts of charge

that is

$$C_g \text{ (MOS)} = C_{base} \text{ (bipolar)}$$

then

$$g_m \text{ (bipolar)} \gg g_m \text{ (MOS)}$$

noting that

$$\begin{aligned} C_{base} &= \tau_F I_c (q/kT) \\ C_g &= C_0 A \end{aligned}$$

where  $C_0$  (often denoted as  $C_{ox}$ ) is the gate to channel capacitance per unit area and  $A = WL$ .  $\tau_F$  is the forward transit time.

## 2.12.2 Comparative aspects of key parameters of CMOS and bipolar transistors

In order to put matters in perspective, a comparison of key parameters follows in Table 2-1.

Table 2-1 A comparison of some parameters

| CMOS   | Bipolar                      |
|--|------------------------------|
| $1. I_{ds} = \frac{(\mu C_0) W}{2L} (V_{gs} - V_t)^2$ $= \frac{\beta}{2} (V_{gs} - V_t)^2 \text{ [In saturation]}$ | $I_c = I_s \exp(qV_{be}/kT)$ |
| $2. g_m = (2\beta)^{1/2} (I_{ds})^{1/2}$ <p>[expressions given can be put in this form]</p>                        | $g_m = I_c / \frac{kT}{q}$   |
| $3. I_{ds}/A = (\mu C_0/2L^2)(V_{gs} - V_t)^2$   | $I_c/A = 1/(R_B \mu \tau_B)$ |

where  $I_{ds}/A$  and  $I_c/A$  are current/area and  $R_B$  is base resistance and  $\tau_B$  is the base transit time (usually in the order of 10–30ps).

Evaluating, we may see that  $I/A$  for bipolar is five times better than that for CMOS. A discussion of the current drive aspects of BiCMOS circuits will be found in Chapter 4 (section 4.8.3).

### 2.12.3 BiCMOS inverters

As in nMOS and CMOS logic circuitry, the basic logic element is the inverter circuit.

When designing with BiCMOS in mind, the logical approach is to use MOS switches to perform the logic function and bipolar transistors to drive the output loads. The simplest logic function is that of inversion, and a simple BiCMOS inverter circuit is readily set out as shown in Figure 2-17.

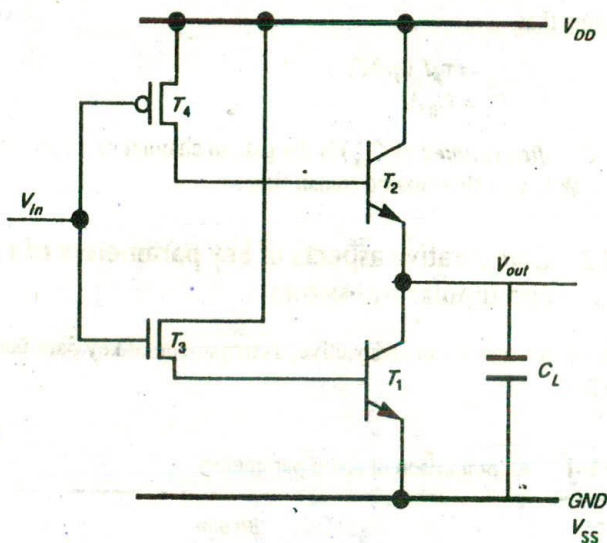


Figure 2-17 A simple BiCMOS Inverter

It consists of two bipolar transistors  $T_1$  and  $T_2$  with one nMOS transistor  $T_3$  and one pMOS transistor  $T_4$ , both being enhancement mode devices. The action of the circuit is straightforward and may be described as follows:

- With  $V_{in} = 0$  volts (GND)  $T_3$  is off so that  $T_1$  will be non-conducting. But  $T_4$  is on and supplies current to the base of  $T_2$  which will conduct and act as a current source to charge the load  $C_L$  toward +5 volts ( $V_{DD}$ ). The output of the inverter will rise to +5 volts less the base to emitter voltage  $V_{BE}$  of  $T_2$ .
- With  $V_{in} = +5$  volts ( $V_{DD}$ )  $T_4$  is off so that  $T_2$  will be non-conducting. But  $T_3$  will now be on and will supply current to the base of  $T_1$  which will conduct and act as a current sink to the load  $C_L$  discharging it toward 0 volts (GND). The output of the inverter will fall to 0 volts plus the saturation voltage  $V_{CEsat}$  from the collector to the emitter of  $T_1$ .
- $T_1$  and  $T_2$  will present low impedances when turned on into saturation and the load  $C_L$  will be charged or discharged rapidly.

- The output logic levels will be good and will be close to the rail voltages since  $V_{CEsat}$  is quite small and  $V_{BE}$  is approximately +0.7 volts.
- The inverter has a high input impedance.
- The inverter has a low output impedance.
- The inverter has a high current drive capability but occupies a relatively small area.
- The inverter has high noise margins.

However, owing to the presence of a DC path from  $V_{DD}$  to  $GND$  through  $T_3$  and  $T_1$ , this is not a good arrangement to implement since there will be a significant static current flow whenever  $V_{in} = \text{logic 1}$ . There is also a problem in that there is no discharge path for current from the base of either bipolar transistor when it is being turned off. This will slow down the action of this circuit.

An improved version of this circuit is given in Figure 2-18, in which the DC path through  $T_3$  and  $T_1$  is eliminated, but the output voltage swing is now reduced, since the output cannot fall below the base to emitter voltage  $V_{BE}$  of  $T_1$ .

An improved inverter arrangement, using resistors, is shown in Figure 2-19. In this circuit resistors provide the improved swing of output voltage when each bipolar transistor is off, and also provide discharge paths for base current during turn-off.

The provision of on chip resistors of suitable value is not always convenient and may be space-consuming, so that other arrangements — such as in Figure 2-20 — are used. In this circuit, the transistors  $T_3$  and  $T_6$  are arranged to turn on when  $T_2$  and  $T_1$  respectively are being turned off.

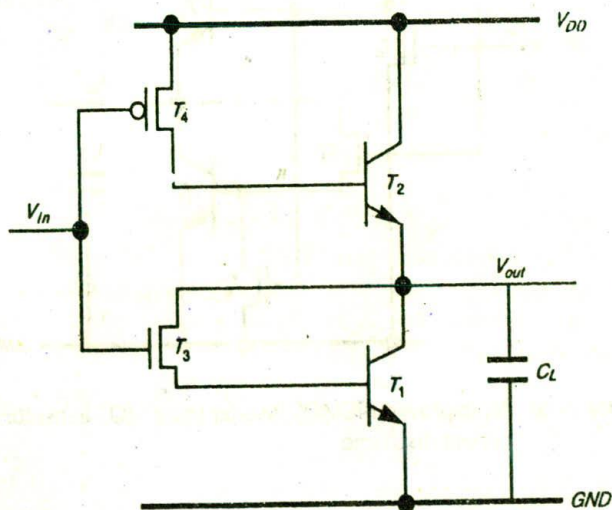


Figure 2-18 An alternative BiCMOS inverter with no static current flow



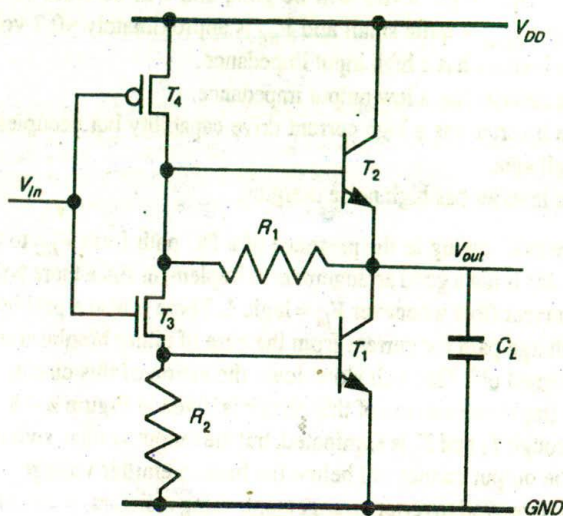


Figure 2-19 An improved BiCMOS inverter with better output logic levels

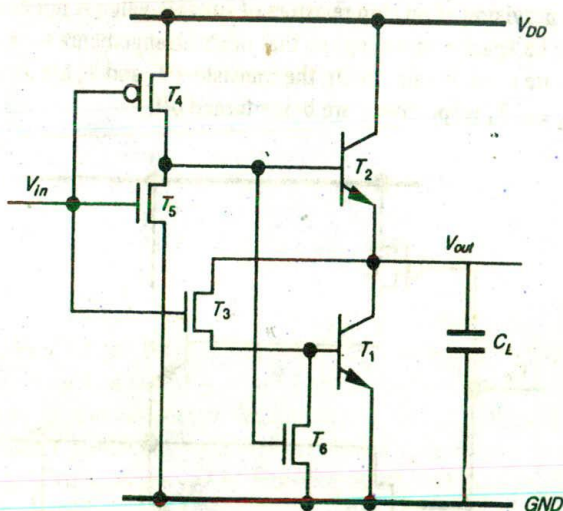


Figure 2-20 An improved BiCMOS inverter using MOS transistors for base current discharge

In general, BiCMOS inverters offer many advantages where high load current sinking and sourcing is required. The arrangements lead on to the BiCMOS gate circuits which will be dealt with in Chapter 5.

## 2.13 Latch-up in CMOS circuits

A problem which is inherent in the p-well and n-well processes is due to the relatively large number of junctions which are formed in these structures and, as mentioned earlier, the consequent presence of parasitic transistors and diodes. Latch-up is a condition in which the parasitic components give rise to the establishment of low-resistance conducting paths between  $V_{DD}$  and  $V_{SS}$  with disastrous results. Careful control during fabrication is necessary to avoid this problem.

Latch-up may be induced by glitches on the supply rails or by incident radiation. The mechanism involved may be understood by referring to Figure 2-21, which shows the key parasitic components associated with a p-well structure in which an inverter circuit (for example) has been formed.

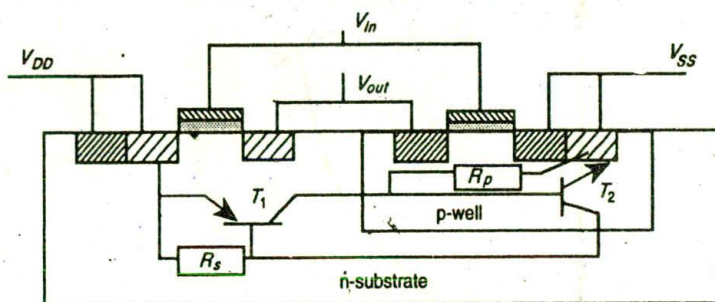


Figure 2-21 Latch-up effect in p-well structure

There are, in effect, two transistors and two resistances (associated with the p-well and with regions of the substrate) which form a path between  $V_{DD}$  and  $V_{SS}$ . If sufficient substrate current flows to generate enough voltage across  $R_s$  to turn on transistor  $T_1$ , this will then draw current through  $R_p$  and, if the voltage developed is sufficient,  $T_2$  will also turn on, establishing a self-sustaining low-resistance path between the supply rails. If the current gains of the two transistors are such that  $\beta_1 \times \beta_2 > 1$ , latch-up may occur. Equivalent circuits are given in Figure 2-22.

With no injected current, the parasitic transistors will exhibit high resistance, but sufficient substrate current flow will cause switching to the low-resistance state as already explained. The switching characteristic of the arrangement is outlined in Figure 2-23.

Once latched-up, this condition will be maintained until the latch-up current drops below  $I_L$ . It is thus essential for a CMOS process to ensure that  $V_L$  and  $I_L$  are not readily achieved in any normal mode of operation.

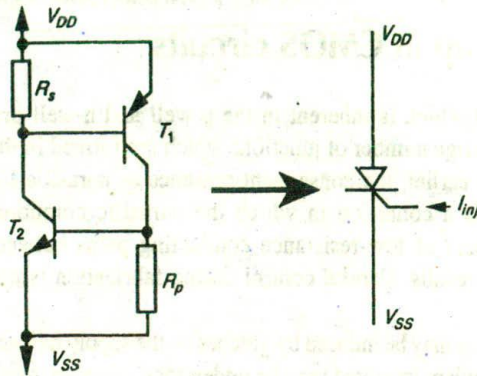


Figure 2-22 Latch-up circuit model

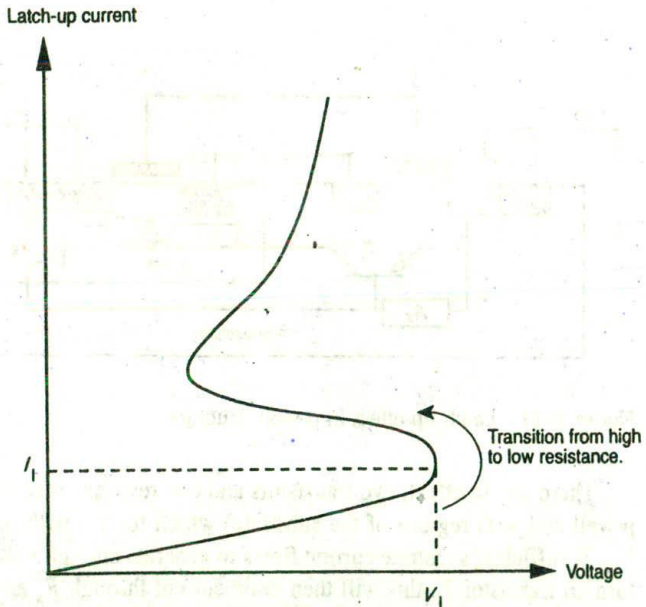


Figure 2-23 Latch-up current versus voltage

Remedies for the latch-up problem include:

1. an increase in substrate doping levels with a consequent drop in the value of  $R_s$ ;
2. reducing  $R_p$  by control of fabrication parameters and by ensuring a low contact resistance to  $V_{SS}$ ;
3. other more elaborate measures such as the introduction of guard rings.



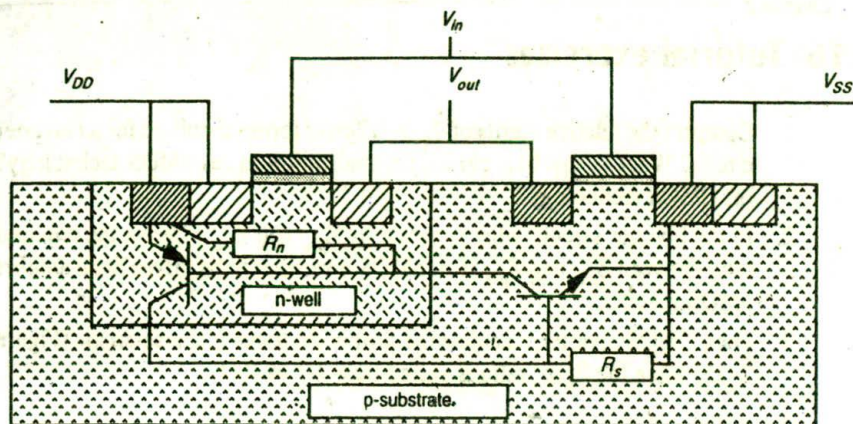


Figure 2-24 Latch-up circuit for n-well process

For completeness, the latch-up configuration for an n-well structure is given in Figure 2-24.

## 2.14 BiCMOS Latch-up susceptibility

One benefit of the BiCMOS process is that it produces circuits which are less likely to suffer from latch-up problems. This is due to several factors:

- A reduction of substrate resistance  $R_s$ .
- A reduction of n-well resistance  $R_w$ .
- A reduction of  $R_s$  and  $R_w$  means that a larger lateral current is necessary to invite latch-up and a higher value of holding current is also required.
- The parasitic (vertical) pnp transistor which is part of the n-well latch-up circuit has its beta reduced owing to the presence of the buried n+ layer. This has the effect of reducing carrier lifetime in the n-base region and this contributes the reduction in beta.

## 2.15 Observations

This chapter has established the underlying properties of MOS active devices and simple circuits configured when using them. The reason for such encumbrances as ratio rules has been explained and it is now appropriate to discuss the means by which circuits can be interconnected in silicon.

## 2.16 Tutorial exercises

1. Compare the relative merits of three different forms of pull-up for an inverter circuit. What is the best choice for realization in (a) nMOS technology? (b) CMOS technology?
2. In the inverter circuit, what is meant by  $Z_{p.u.}$  and  $Z_{p.d.}$ ? Derive the required ratio between  $Z_{p.u.}$  and  $Z_{p.d.}$  if an nMOS inverter is to be driven from another nMOS inverter.
3. For a CMOS inverter, calculate the shift in the transfer characteristic (Figure 2-15) when the  $\beta_n/\beta_p$  ratio is varied from 1/1 to 10/1.



# MOS and BiCMOS circuit design processes

*The artist must understand that he does not (only) create — he materializes.*

Horia Bernea

## Objectives

The purpose of this chapter is to provide an insight into the methods and means for materializing circuit designs in silicon.

Design processes are aided by simple concepts such as stick and symbolic diagrams but the key element is a set of design rules. Design rules are the communication link between the designer specifying requirements and the fabricator who materializes them. Design rules are used to produce workable mask layouts from which the various layers in silicon will be formed or patterned.

The first set of design rules introduced here are 'lambda-based'. These rules are straightforward and relatively simple to apply. However, they are 'real' and chips can be fabricated from mask layouts using the lambda-based rule set.

Tighter and faster designs will be realized if a fabricator's line is used to its full advantage and such rule sets are generally particular not only to the fabricator but also to a specific technology.

Two such design rule sets, from Orbit\*, are also introduced in this chapter.

---

\*Orbit Semiconductor Inc., California.



## 3.1 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-diffusion*, *polysilicon*, and *metal*, which are isolated from one another by thick or thin (thin oxide) silicon dioxide insulating layers. The thin oxide (thin oxide) mask region includes *n-diffusion*, *p-diffusion*, and transistor channels. Polysilicon and thin oxide 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 be deliberately joined together where contacts are formed. We have also seen that the basic MOS transistor properties can be modified by the use of an implant within the thin oxide region and this is used in nMOS circuits to produce depletion mode transistors.

We have also seen that bipolar transistors can be included in this design process by the addition of extra layers to a CMOS process. This is referred to as BiCMOS technology, and in this text it is dealt with in an n-well CMOS environment.

We must find a way of capturing the topology and layer information of the actual circuit in silicon so that we can set out simple diagrams which convey both *layer* information and *topology*.

## 3.2 Stick diagrams

Stick diagrams may be used to convey layer information through the use of a color code — for 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. In this text the color coding has been complemented by monochrome encoding of the lines so that black and white copies of stick diagrams do not lose the layer information. The encodings chosen are shown and illustrated in color as Color plates 1(a)–(d) and in monochrome form as Figures 3–1(a)–(d). When you are drawing your own stick diagrams you should use single lines in the appropriate colors, as in Color plate 1(d) noting that yellow lines are outlined in green for clarity only.

Note that mask layout information, which is also color coded, may also be hatched for monochrome encoding, also shown in Figures 3–1(a)–(c). Monochrome encoding schemes are widely illustrated throughout the text, and it will be noted that diagrams and mask layouts in this form are readily reproduced by copying machines.

The color and monochrome encoding scheme used has been evolved to cover nMOS, CMOS, and BiCMOS processes and to be compatible with the design processes of gallium arsenide. The color encoding is compatible with color terminals, printers, and plotters having quite simple color palettes. Using color workstations,

| COLOR  | STICK ENCODING                  | LAYERS   | MASK LAYOUT ENCODING           | CIF LAYER |
|--|---------------------------------|--|--------------------------------|-----------|
|  | MONOCHROME                      |  | MONOCHROME                     |           |
| GREEN  |                                 | n-diffusion<br>(n <sup>+</sup> active)<br>Thinox * |                                | ND        |
| RED  |                                 | Polysilicon  |                                | NP        |
| BLUE   |                                 | Metal 1  |                                | NM        |
| BLACK  |                                 | Contact cut  |                                | NC        |
| GRAY   | NOT APPLICABLE                  | Overglass  |                                | NG        |
| nMOS ONLY<br>YELLOW  |                                 | Implant  |                                | NI        |
| nMOS ONLY<br>BROWN   |                                 | Buried contact                                     |                                | NB        |
| FEATURE  | FEATURE (STICK)<br>(MONOCHROME) | FEATURE (SYMBOL)<br>(MONOCHROME)                   | FEATURE (MASK)<br>(MONOCHROME) |           |
| n-type enhancement mode transistor   |                                 |  |                                |           |
| Transistor length to width ratio L:W should be shown but source, drain and gate labeling will not normally be shown. |                                 |  |                                |           |
| n-type depletion mode transistor<br>nMOS ONLY  |                                 |  |                                |           |

Figure 3-1(a) Encodings for a simple single metal nMOS process (see Color plate 1(a) for nMOS color encoding details)

the mask areas are usually color filled while pen plotters produce color outlines only. In this text, most color diagrams incorporate color outlines and color hatching (hatching as for the monochrome encoding) so that the detail of underlying areas may be easily discerned where layers intersect or are superimposed. This form of color representation is acceptable for those with color vision difficulties and may also be copied by a monochrome copier without losing the encoding. The various representations are indicated in Color plate 2.




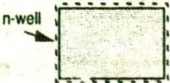


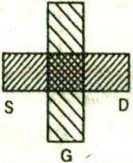







| COLOR   | STICK ENCODING  | LAYERS  | MASK LAYOUT ENCODING         | CIF LAYER  |
|---|---|---|------------------------------|------------|
|   | MONOCHROME  |   | MONOCHROME                   |            |
| GREEN   | ENCODING AS IN FIGURE 3-1(a)  | n-diffusion (n <sup>+</sup> active)<br>Thin <sub>ox</sub> * | ENCODING AS IN FIGURE 3-1(a) | CAA or CNA |
| RED   |   | Polysilicon   |                              | CPF        |
| BLUE  |   | Metal 1   |                              | CMF        |
| BLACK   |   | Contact cut   |                              | CC         |
| GRAY  |   | Overglass   |                              | COG        |
| GREEN IN P <sup>+</sup> (MASK)                              |   | p-diffusion (p <sup>+</sup> active)                         |                              | CAA or CPA |
| YELLOW (STICK)  | NOT SHOWN IN STICK DIAGRAM  | p <sup>+</sup> mask   |                              | CPP        |
| YELLOW  |   |   |                              |            |
| DARK BLUE OR PURPLE   |   | Metal 2   |                              | CMS        |
| BLACK   |   | VIA   |                              | CVA        |
| BROWN   | DEMARICATION LINE<br>p-well edge is shown as a demarcation line in stick diagrams | p-well  |                              | CPW        |
| BLACK   |   | V <sub>DD</sub> or V <sub>SS</sub> CONTACT                  |                              | CC         |
| FEATURE   | FEATURE (STICK) (MONOCHROME)  | FEATURE (SYMBOL) (MONOCHROME)                               | FEATURE (MASK) (MONOCHROME)  |            |
| n-type enhancement mode transistor<br>(as in Figure 3-1(a)) |   |   |                              |            |
| p-type enhancement mode transistor                          |   |   |                              |            |

The same well encoding and demarcation line are used for an n-well process.  
 For p-well process, the n features are in the well. For an n-well process, the p features are in the well.

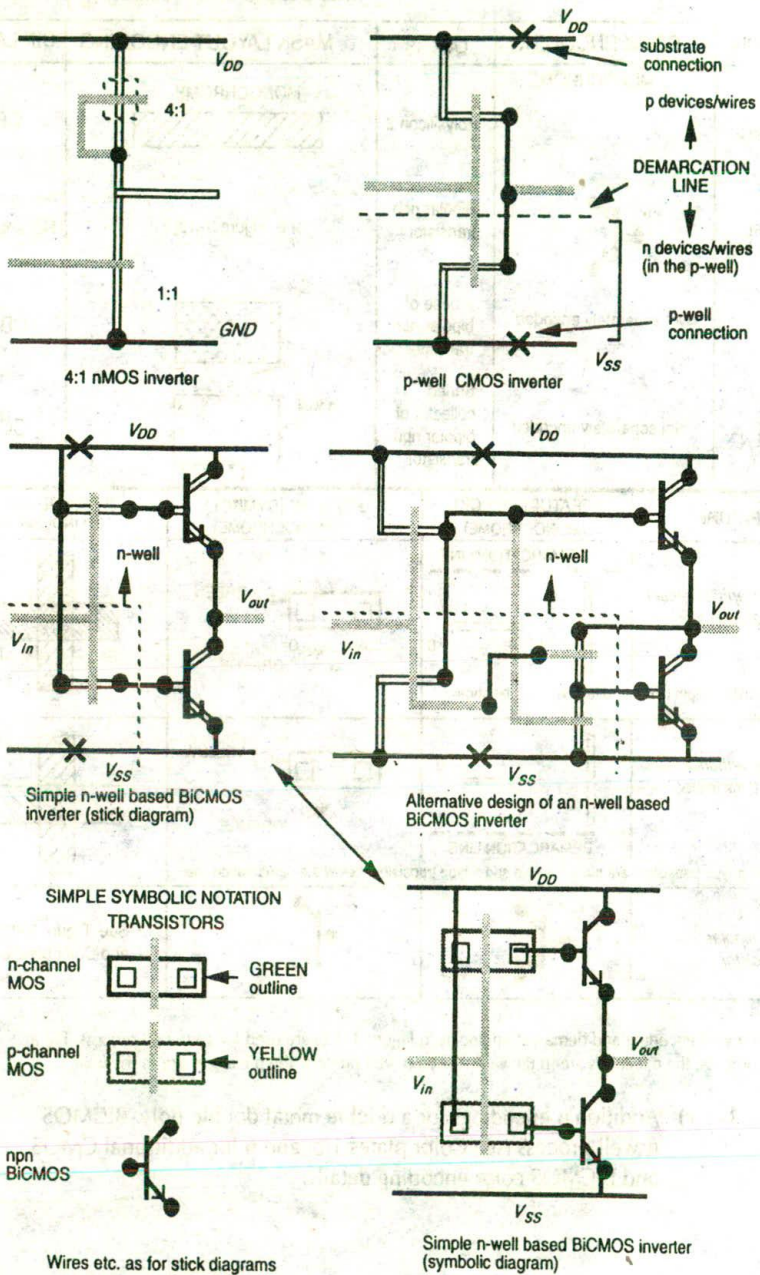
Figure 3-1(b) Encodings for a double metal CMOS p-well process (see Color plate 1(b) for CMOS color encoding details)



| COLOR   | STICK ENCODING   | LAYERS  | MASK LAYOUT ENCODING  | CIF LAYER      |
|---|--|---|---|----------------|
| ORANGE  | MONOCHROME   | Polysilicon 2   | MONOCHROME<br> | CPS            |
| SEE COLOR PLATE 1(c)                          |   | Bipolar npn transistor  | see Figure 3-13(f)  | Not applicable |
| PINK  | Not separately encoded   | p-base of bipolar npn transistor  |                | CBA            |
| PALE GREEN                                    | Not separately encoded   | Buried collector of bipolar npn transistor  | n-well<br>     | CCA            |
| FEATURE                                       | FEATURE (STICK) (MONOCHROME)   | FEATURE (SYMBOL) (MONOCHROME)   | FEATURE (MASK) (MONOCHROME)   |                |
| <i>n</i> -type enhancement poly. 2 transistor | DEMARCATION LINE<br><br>L:W<br>S G D<br>Transistor length to width ratio L:W may be shown.                                  | <br>GREEN<br>S G D<br>ORANGE | <br>S D<br>G   |                |
| <i>p</i> -type enhancement poly. 2 transistor | <br>L:W<br>DEMARCATION LINE<br>Note: p-type transistors are placed above and n-type transistors below the demarcation line. | <br>YELLOW<br>ORANGE         |                |                |
| <i>npn bipolar transistor</i>                 |   |                            | See Figure 3-13(f) and Color plate 6  |                |

The same well encoding and demarcation line as in Figure 3-1(b) are used for an n-well process. For a p-well process, the n features are in the well. For an n-well process, the p features are in the well.

**Figure 3-1(c)** Additional encodings for a double metal double poly. BiCMOS n-well process (see Color plates 1(c) and 6 for additional CMOS and BiCMOS color encoding details)



Monochrome stick diagram examples

Figure 3-1(d) Stick diagrams and simple symbolic encoding (see also Color plate 1(d))



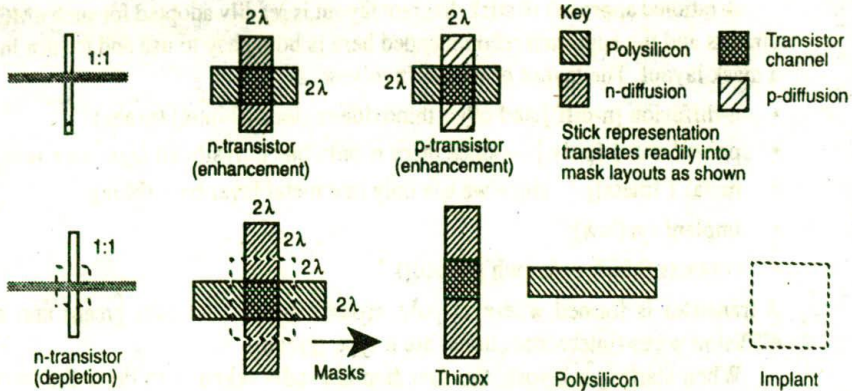


Figure 3-2 Stick diagrams and corresponding mask layout examples

In order to facilitate the learning and use of the encoding schemes, the simple set required for a single metal nMOS design is set out first as Figure 3-1(a) and Color plate 1(a); for a double metal CMOS p-well process the required encodings are extended by those given as Figure 3-1(b) and Color plate 1(b). Figure 3-1(c) and Color plate 1(c) further extend the representations to cover a second polysilicon layer and BiCMOS technology.

In this chapter we will see how basic circuits are represented in stick diagram and in symbolic form. We will be using stick representation quite widely throughout the text. The layout of stick diagrams faithfully reflects the topology of the actual layout in silicon. To illustrate stick diagrams, inverter circuits are presented in Figure 3-1(d) and in Color plate 1(d) — in nMOS, in p-well CMOS, and in n-well BiCMOS technology. A symbolic form of diagram is often most convenient and such diagrams are based on the simple symbol set included in Figures 3-1(a)–(c) and Color plates 1(a)–(c). The simplicity of symbolic form is illustrated in Figure 3-1(d), in Color plate 1(d), and in Color plate 7.

Having conveyed layer information and topology by using stick or symbolic diagrams, these diagrams are relatively easily turned into mask layouts as, for example, the transistor stick diagrams of Figure 3-2 stressing the ready translation into mask layout form.

In order that the mask layouts produced during design will be compatible with the fabrication processes, a set of design rules are set out for layouts so that, if obeyed, the rules will produce layouts which will work in practice.

### 3.2.1 nMOS design style

In order to start with a relatively simple process, we will consider single metal, single polysilicon nMOS technology (see Figure 3-1(a) and Color plate 1(a)).



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 turn into a mask layout. The layout of nMOS involves:

- n-diffusion [n-diff.] and other thinoxide regions [thinox] (green);
- polysilicon 1 [poly.] — since there is only one polysilicon layer here (red);
- metal 1 [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)  $V_{DD}$  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 3-3(a), 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  $V_{DD}$  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 3-3(b), 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-like circuits.

Signal paths may also be switched by pass transistors, and long signal paths may often require metal buses (blue). Allowing for the fact that the stick diagram may well represent only a small section of circuit which will be replicated many times, a convenient strategy is to run power rails and bus(es) in parallel in metal (blue) and then propagate control signals at right angles on poly. as shown. At this stage of design, 'leaf-cell' boundaries are conveniently shown on the stick diagram and these are placed so that replicated cells may be directly interconnected by direct abutment on a side-by-side and/or top-to-bottom basis. The aspects just discussed are illustrated in Figure 3-3(c).

From the very beginning a design style should encourage the concepts of 'regularity' (through the use of replication) and generality so that design effort can be minimized and the interconnection of leaf-cells, subsystems and systems is facilitated.

### 3.2.2 CMOS design style

The stick and layout representations for CMOS used in this text are a logical extension of the nMOS approach and style already outlined. They are based on the widely accepted work of Mead and Conway.

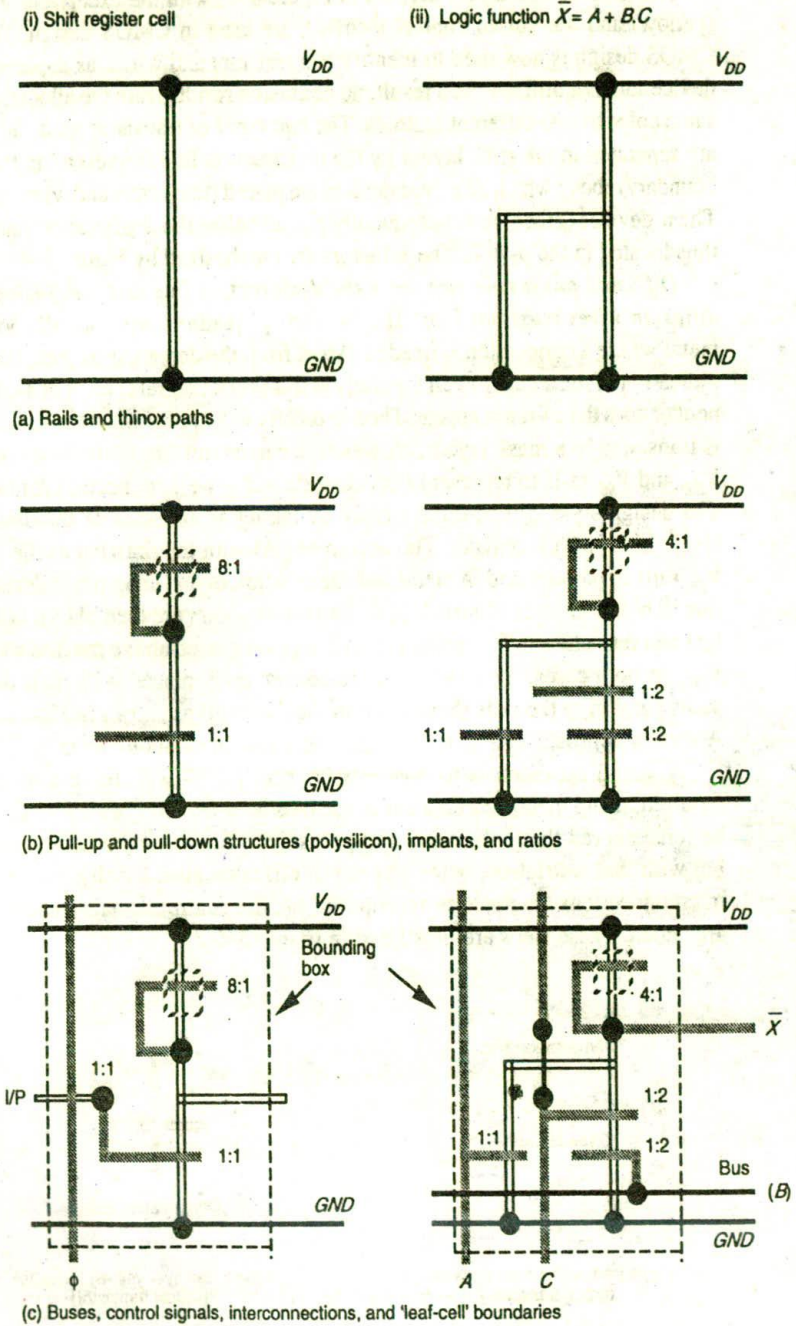


Figure 3-3 Examples of nMOS stick layout design style



All features and layers defined in Figure 3-1, 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. These factors are emphasized by Figure 3-4.

*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. Apart from the demarcation line, there is no indication of the actual p-well topology at this (stick diagram) level of abstraction; neither does the p+ mask appear. Their geometry will appear when the stick diagram is translated to a mask layout. However, we must not forget to place crosses on  $V_{DD}$  and  $V_{SS}$  rails to represent the substrate and p-well connection respectively. The design style is illustrated simply by taking as an example the design of a single bit of a shift register. The design begins with the drawing of the  $V_{DD}$  and  $V_{SS}$  rails in parallel and in metal and the creation of an (imaginary) demarcation line in between, as in Figure 3-5(a). The n-transistors are then placed below this line and thus close to  $V_{SS}$ , while p-transistors are placed above the line and below  $V_{DD}$ . In both cases, the transistors are conveniently placed with their diffusion paths parallel to the rails (horizontal in the diagram) as shown in Figure 3-5(b). A similar approach can be taken with transistors in symbolic form.

A sound approach is to now interconnect the n- with the p-transistors as required, using metal and connect to the rails as shown in Figure 3-5(c). It must be remembered that only metal and polysilicon can cross the demarcation line but with that restriction, wires can run in diffusion also. Finally, the remaining interconnections are made as appropriate and the control signals and data inputs are added. These steps are illustrated in Figure 3-5(d).

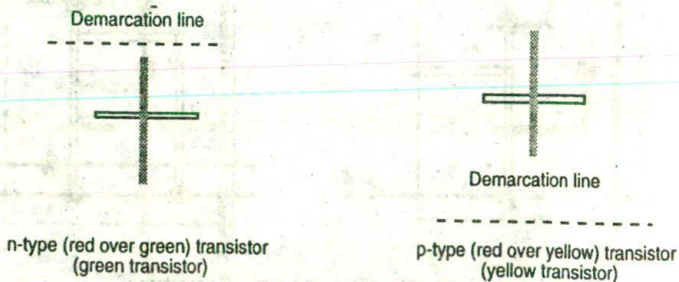
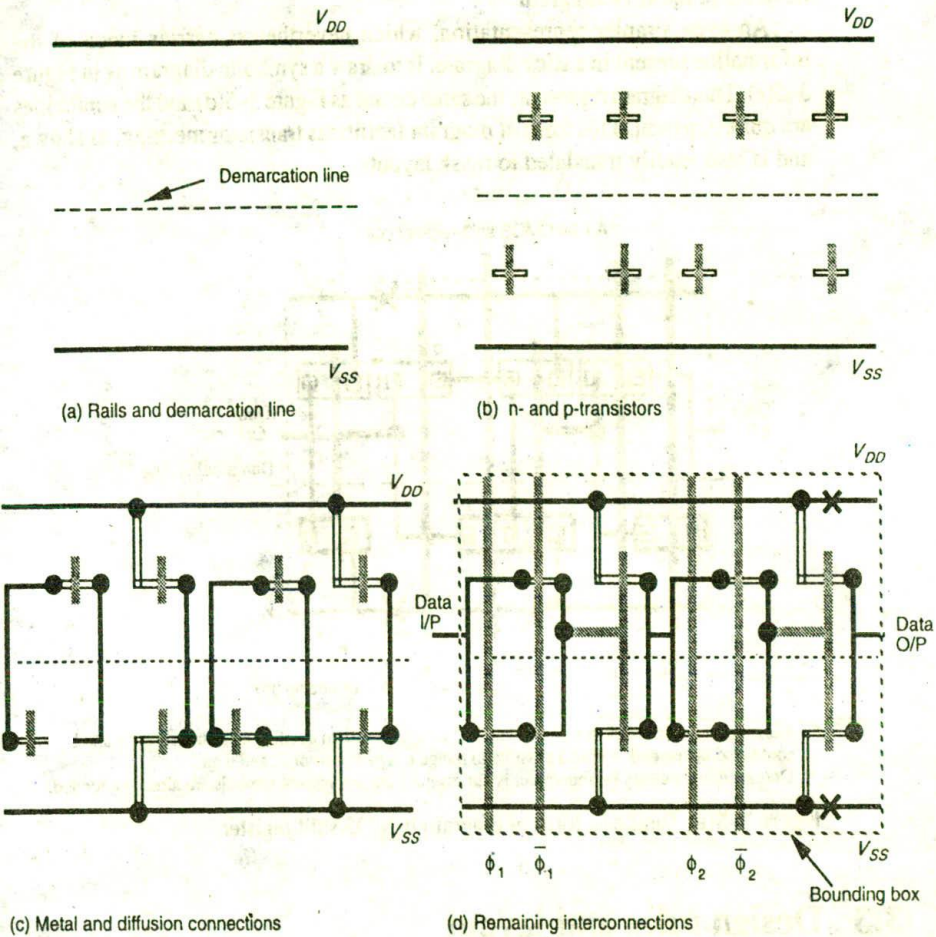


Figure 3-4 n-type and p-type transistors in CMOS design



(Using a 1-bit shift register stage as an example)



Note: The contact crosses in (d) should represent one  $V_{DD}$  contact for every four p-transistors and one  $V_{SS}$  contact for every four n-transistors.

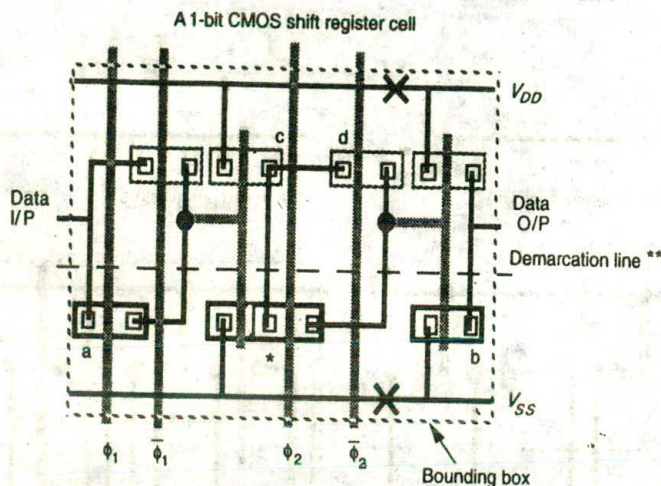
Figure 3-5 Example of CMOS stick layout design style

Although the circuit layout is now complete, we must not forget to represent the  $V_{SS}$  and  $V_{DD}$  contact crosses — one on the  $V_{DD}$  line for every four p-transistors and one on the  $V_{SS}$  line for every four n-transistors. The bounding box for the entire leaf-cell may also be shown if appropriate.

This design style is straightforward in application but later on we may recognize that sometimes transistors can be merged to advantage. We will also see how

stick diagrams are turned into mask layouts, noting for CMOS layouts that the thinox mask includes all green features (n-devices) and all yellow features (p-devices) in the stick diagram.

An even simpler representation, which nevertheless carries much of the information present in a stick diagram, is to draw a symbolic diagram as in Figure 3-5(e). This diagram represents the same circuit as Figure 3-5(d) and the similarities are quite apparent. This form of diagram facilitates transistor merging, as shown, and is also readily translated to mask layouts.



- \* Note that two transistors (n-type) are merged as shown. When abutting cells, transistors a and b could also be merged. It is also possible to merge p-type transistors c and d etc.
- \*\* Demarcation line may be shown but is not essential since transistor symbols are already encoded.

Figure 3-5(e) Symbolic form of diagram (CMOS shift register)

### 3.3 Design rules and layout

The object of a set of design rules is to allow a ready translation of circuit design concepts, usually in stick diagram or symbolic form, into actual geometry in silicon. The design rules are the effective interface between the circuit/system designer and the fabrication engineer. Clearly, both sides of the interface have a vested interest in making their own particular tasks as easy as possible and design rules usually attempt to provide a workable and reliable compromise that is friendly to both sides.

Circuit designers in general want tighter, smaller layouts for improved performance and decreased silicon area. On the other hand, the process engineer wants design rules that result in a *controllable and reproducible* process. Generally



we find that there has to be a compromise for a competitive circuit to be produced at a reasonable cost.

One of the important factors associated with design rules is the achievable definition of the process line. Definition is determined by process line equipment and process design. For example, it is found that if a 10:1 wafer stepper is used instead of a 1:1 projection mask aligner, the level-to-level registration will be closer. Design rules can be affected by the maturity of the process line. For example, if the process is mature, then one can be assured of the process line capability, allowing tighter designs with fewer constraints on the designer.

The simple ' $\lambda$ -based' design rules set out first in this text are based on the invaluable work of Mead and Conway and have been widely used, particularly in the educational context and in the design of multiproject chips. The design rules are based on a single parameter  $\lambda$  which leads to a simple set of rules for the designer, and wide acceptance of the rules by a large cross-section of the fabrication houses and silicon brokers, and allows for scaling of the designs to a limited extent. This latter feature may help to give designs a longer lifetime. The simplicity of  $\lambda$ -based rules also provides a simple introduction to design rules and to mask layout design in general and helps to set the scene for the 'micron-based' rule sets which follow.

### 3.3.1 Lambda-based design rules

In general, design rules and layout methodology based on the concept of  $\lambda$  provide a process and feature size-independent way of setting out mask dimensions to scale.

All paths in all layers will be dimensioned in  $\lambda$  units and subsequently  $\lambda$  can be allocated an appropriate value compatible with the feature size of the fabrication process. This concept means that the actual mask layout design takes little account of the value subsequently allocated to the feature size, but the design rules are such that, if correctly obeyed, the mask layouts will produce working circuits for a range of values allocated to  $\lambda$ . For example,  $\lambda$  can be allocated a value of 1.0  $\mu\text{m}$  so that minimum feature size on chip will be 2  $\mu\text{m}$  ( $2\lambda$ ). Design rules, also due to Mead and Conway, specify line widths, separations, and extensions in terms of  $\lambda$ , and are readily committed to memory. Design rules can be conveniently set out in diagrammatic form as in Figure 3-6 for the widths and separation of conducting paths, and in Figure 3-7 for extensions and separations associated with transistor layouts.

The design rules associated with contacts between layers are set out in Figures 3-8 and 3-9 and it will be noted that connection can be made between two or, in the case of nMOS designs, three layers.

In all cases, the use of the design rules will be illustrated in layouts resulting from exercises worked through in the text.



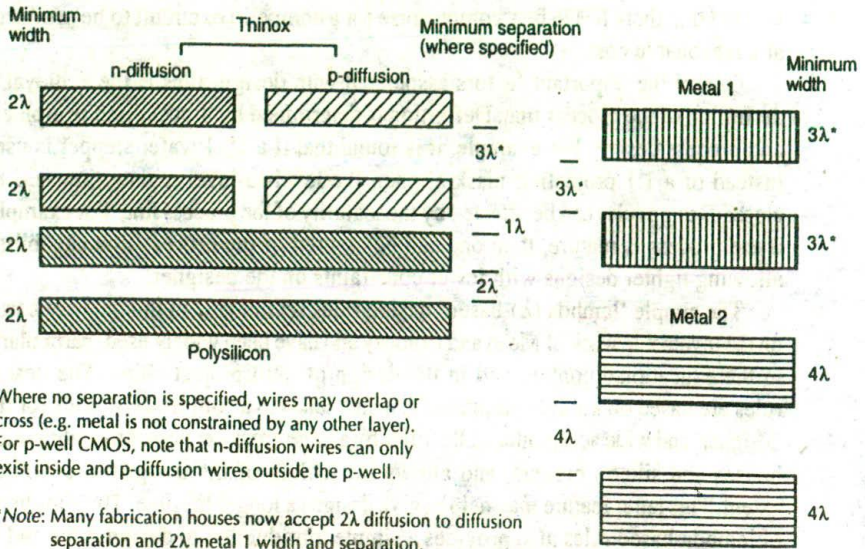


Figure 3-6 Design rules for wires (nMOS and CMOS)

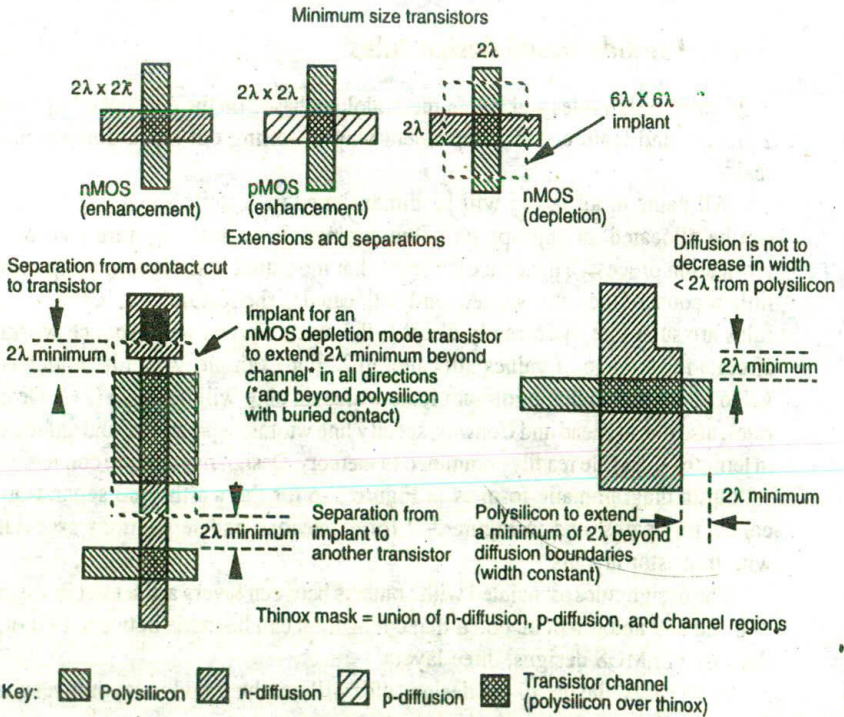
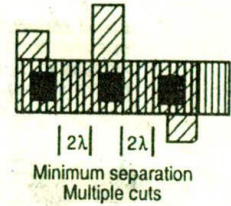


Figure 3-7 Transistor design rules (nMOS, pMOS, and CMOS)

## 1. Metal 1 to polysilicon or to diffusion

3 $\lambda$  minimum

2 $\lambda$  x 2 $\lambda$  cut centered  
on 4 $\lambda$  x 4 $\lambda$  superimposed  
areas of layers to be joined  
in all cases



## 2. Via (contact from metal 2 to metal 1 and thence to other layers)

Metal 2



2 $\lambda$  minimum separation  
(if other spacings allow)

4 $\lambda$  x 4 $\lambda$  area of overlap with  
2 $\lambda$  x 2 $\lambda$  via at center

Via and cut used to  
connect metal 2 to  
diffusion



Figure 3-8 Contacts (nMOS and CMOS)

## 3.3.2 Contact cuts

When making contacts between polysilicon and diffusion in nMOS circuits it should be recognized that there are three possible approaches — poly. to metal then metal to diff., or a *buried contact* poly. to diff., or a *butting contact* (poly. to diff. using metal). Of the latter two, the buried contact is the most widely used, giving economy in space and a reliable contact. Butting contacts were widely used at one time but have been mostly superseded by buried contacts and have been included here and in the figures for the sake of completeness. In CMOS designs, poly. to diff. contacts are almost always made via metal.

When making connections between metal and either of the other two layers (as in Figure 3-8), the process is quite simple. The 2 $\lambda$  x 2 $\lambda$  contact cut indicates an area in which the oxide is to be removed down to the underlying polysilicon or diffusion surface. When deposition of the metal layer takes place the metal is deposited through the contact cut areas onto the underlying area so that contact is made between the layers.

When connecting diffusion to polysilicon using the butting contact approach (see Figure 3-9), the process is rather more complex. In effect, a 2 $\lambda$  x 2 $\lambda$  contact cut is made down to each of the layers to be joined. The layers are butted together in such a way that these two contact cuts become contiguous. Since the polysilicon and diffusion outlines overlap and thin oxide under polysilicon acts as a mask in



**Buried contact:** Basically, layers are joined over a  $2\lambda \times 2\lambda$  area with the buried contact cut extending by  $1\lambda$  in all directions around the contact area except that the contact cut extension is increased to  $2\lambda$  in diffusion paths leaving the contact area. This is to avoid forming unwanted transistors (see following examples).

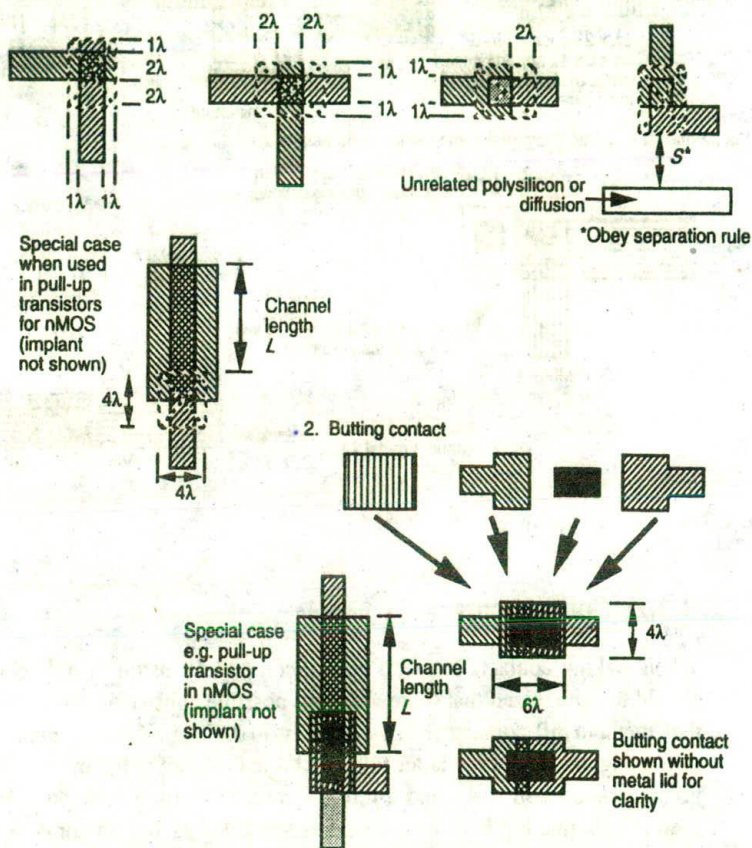


Figure 3-9 Contacts polysilicon to diffusion (nMOS only in the main text)

the diffusion process, the polysilicon and diffusion layers are also butted together. The contact between the two butting layers is then made by a metal overlay as shown in the figure. It is hoped that the cross-sectional view of the butting contact in Figure 3-10(b) helps to make the nature of the contact apparent.

The buried contact approach shown in Figures 3-9 and 3-10 is simpler, the contact cut (broken line) in this case indicating where the thin oxide is to be removed to reveal the surface of the silicon wafer before polysilicon is deposited. Thus, the polysilicon is deposited directly on the underlying crystalline wafer. When diffusion takes place, impurities will diffuse into the polysilicon as well as

into the diffusion region within the contact area. Thus a satisfactory connection between polysilicon and diffusion is ensured. Buried contacts can be smaller in area than their butting contact counterparts and, since they use no metal layer, they are subject to fewer design rule restrictions in a layout.

The design rules in this case ensure that a reasonable contact area is achieved and that there will be no transistor formed unintentionally in series with the contact. The rules are such that they also avoid the formation of unwanted diffusion to polysilicon contacts and protect the gate oxide of any transistors in the vicinity of the buried contact cut area.

### 3.3.3 Double metal MOS process rules

A powerful extension to the process so far described is provided by a second metal layer. This gives a much greater degree of freedom, for example, in distributing global  $V_{DD}$  and  $V_{SS}$  (*GND*) rails in a system. Other processes also allow a second polysilicon layer and one such process will be introduced later.

From the overall chip interconnection aspect, the second metal layer in particular is important and, although the use of such a layer is readily envisaged, its disposition relative to (and details of) its connection to other layers using metal 1 to metal 2 contacts, called *vias*, can be readily established with reference to Figures 3-8 and 3-10(c).

Usually, second level metal layers are coarser than the first (conventional) layer and the isolation layer between the layers may also be of relatively greater thickness. To distinguish contacts between first and second metal layers, they are known as *vias* rather than contact cuts. The second metal layer representation is color coded dark blue (or purple). For the sake of completeness, the process steps for a two-metal layer process are briefly outlined as follows.

The oxide below the first metal layer is deposited by atmospheric chemical vapor deposition (CVD) and the oxide layer between the metal layers is applied in a similar manner. Depending on the process, removal of selected areas of the oxide is accomplished by plasma etching, which is designed to have a high level of vertical ion bombardment to allow for high and uniform etch rates.

Similarly, the bulk of the process steps for a double polysilicon layer process are similar in nature to those already described, except that a second thin oxide layer is grown after depositing and patterning the first polysilicon layer (Poly. 1) to isolate it from the now to be deposited second poly. layer (Poly. 2). The presence of a second poly. layer gives greater flexibility in interconnections and also allows Poly. 2 transistors to be formed by intersecting Poly. 2 and diffusion.

To revert to the double metal process it is convenient at this point to consider the layout strategy commonly used with this process. The approach taken may be summarized as follows:

1. Use the second level metal for the global distribution of power buses, that is,  $V_{DD}$  and *GND* ( $V_{SS}$ ), and for clock lines.



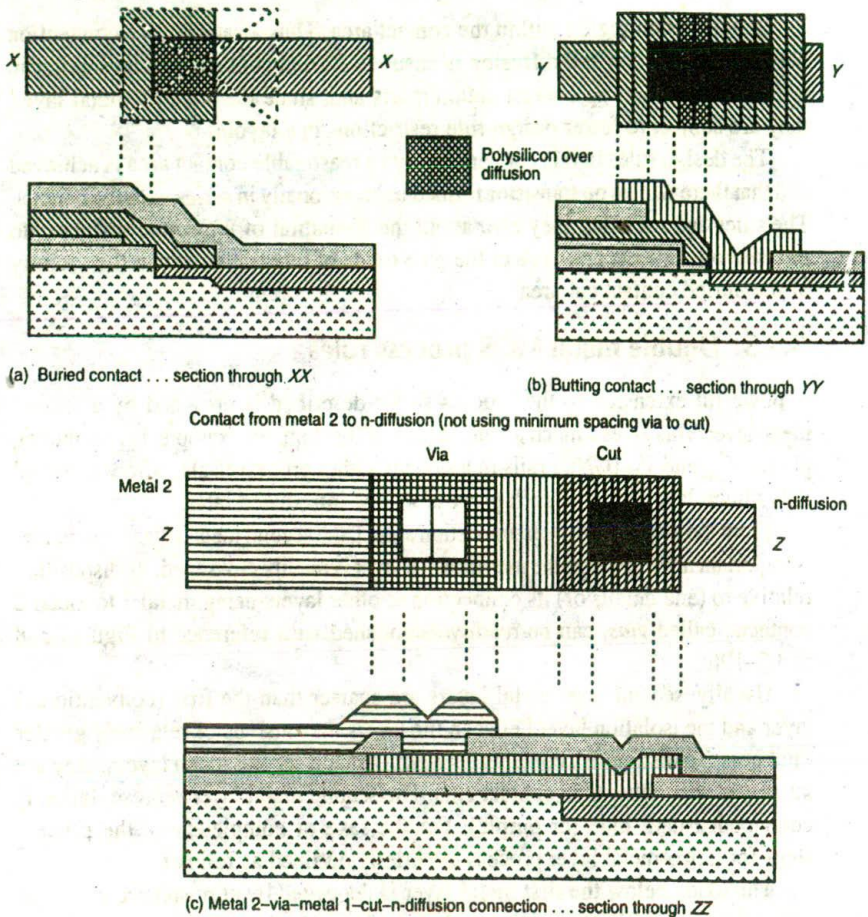


Figure 3-10 Cross-sections through some contact structures

2. Use the first level metal for local distribution of power and for signal lines.
3. Lay out the two metal layers so that the conductors are mutually orthogonal wherever possible.

### 3.3.4 CMOS lambda-based design rules

The CMOS fabrication process is much more complex than nMOS fabrication, which, in turn, has been simplified for ready presentation in this text. The new reader may well think that the design rules discussed here are complex enough,

but in fact they constitute an abstract of the actual processing steps which are used to produce the chip. In a CMOS process, for example, the actual set of industrial design rules may well comprise more than 100 separate rules, the documentation for which spans many pages of text and/or many diagrams. Two such rule sets, micron-based, will be given in this text.

However, extending the Mead and Conway concepts, which we have already set out for pMOS designs, and noting the exclusion of butting and buried contacts, it is possible to add rules peculiar to CMOS (Figure 3-11) to those already set out in Figures 3-6 to 3-10. The additional rules are concerned with those features unique to p-well CMOS, such as the p-well and p+ mask and the special 'substrate' contacts. We have already provided for the p-transistors and p-wires in Figures 3-6 to 3-10. The rules given are also readily translated to an n-well process.

Although the CMOS rules in total may seem difficult to comprehend for the new designer, once use has been made of the simpler nMOS rules the transition to CMOS is not hard to achieve. The real key to success in VLSI design is to put it into practice, and this text attempts to encourage the reader to do just that.

### 3.4 General observations on the design rules

Owing to the microscopic nature of dimensions and features of silicon circuits, a major problem is presented by possible deviation in line widths and in interlayer registration.

If the line widths are too small, it is possible for lines thus defined to be discontinuous in places.

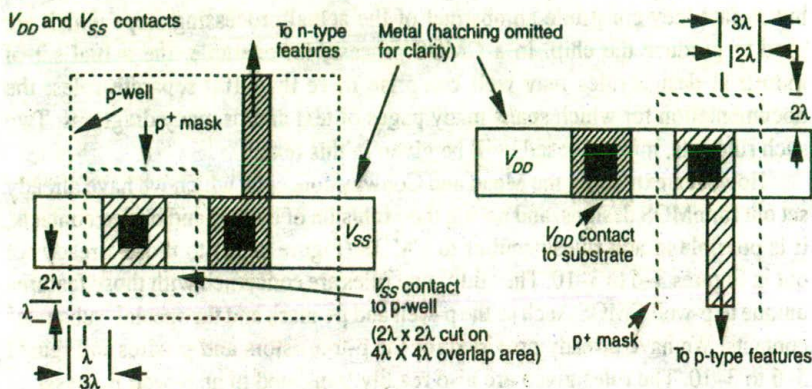
If separate paths in a layer are placed too close together, it is possible that they will merge in places or interfere with each other.

For the lambda-based rules discussed initially, the design rules are formulated in terms of a length unit  $\lambda$  which is related to the resolution of the process.  $\lambda$  may be viewed as a bound on the width deviation of a feature from its ideal 'as drawn' size and also as a bound on the maximum misalignment of any one mask. In the worst case, these effects may combine to cause the relative position of feature edges on different mask levels to deviate by as much as  $2\lambda$  in their interrelationship. Inevitably, a consequence of using the lambda-based concept is that every dimension must be rounded up to whole  $\lambda$  values and this leads to layouts which do not fully exploit the capabilities of the process.

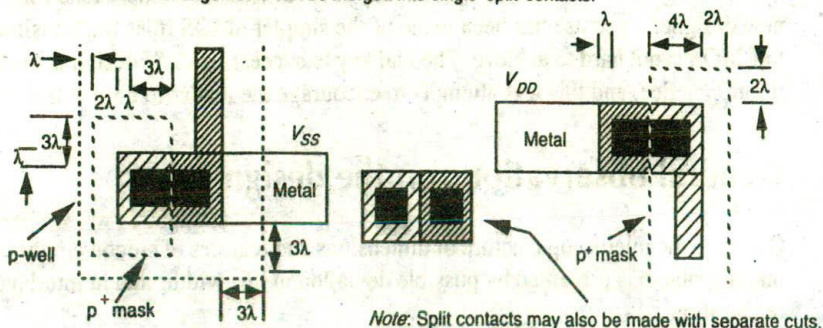
Similar concepts underlie the establishment of 'micron-based' rule sets, but actual dimensions are given so that full advantage can be taken of the fabrication line capabilities and tighter layouts result.

Layout rules, therefore, provide strict guidelines for preparing the geometric layouts which will be used to configure the actual masks used during fabrication and can be regarded as the main communication link between circuit/systems designers and the process engineers engaged in manufacture.





Each of the above arrangements can be merged into single 'split' contacts.



p-well and  $p^+$  mask rules

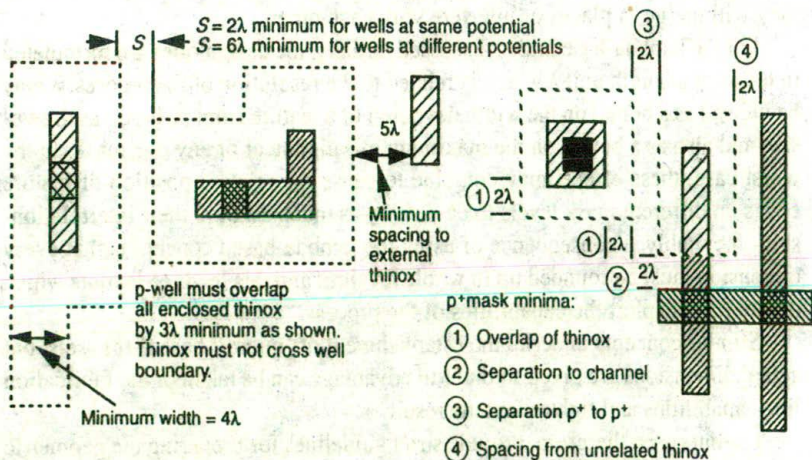


Figure 3-11 Particular rules for p-well CMOS process

The goal of any set of design rules should be to optimize yield while keeping the geometry as small as possible without compromising the reliability of the finished circuit.

On the questions of yield and reliability, even the conservative nature of the lambda-based rules can stand reevaluation when these two factors are of paramount importance. In particular, the rules associated with contacts can be improved upon in the light of experience. Figure 3-12 sets out aspects that may be observed for high yield and in high reliability situations.

In our proposed scheme of events in creating stick layouts for CMOS, we have assumed that poly. and metal can both freely cross well boundaries and this is indeed the case, but we should be careful to try to exclude poly. from areas which lie within p+ mask areas where possible. The reason for this is that the resistance of the poly. layer is reduced in current processes by n-type doping. Clearly the p+ doping which takes place inside the p+ mask will also dope the poly. which is already in place when the p+ doping step takes place. This results in an increase in the n-doping poly. resistance which may be significant in certain parts of a system.

The  $3\lambda$  metal width rule is a conservative one but is implemented to allow for the fact that the metal layer is deposited after the others and on top of them and several layers of silicon dioxide, so that the surface on which it sits is quite 'mountainous'. The metal layer is also light-reflective and these factors combine to result in poor edge definition. In double metal the second layer of metal has an even more uneven terrain on which to be deposited and patterned. Hence metal 2 is often wider than metal 1.

Metal to metal separation is also large and is brought about mainly by difficulties in defining metal edges accurately during masking operations on the highly reflective metal.

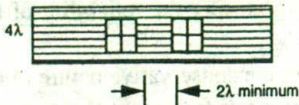
All diffusion processes are such that lateral diffusion occurs as well as impurity penetration from the surface. Hence the separation rules for diffusion allow for this and relatively large separations are specified. This is particularly the case for the p-well diffusions which are deep diffusions and thus have considerable lateral spread.

Transitions from thin gate oxide to thick field oxide in the oxidation process also use up space and this is another reason why the lambda-based rules require a minimum separation between thinox regions of  $3\lambda$ . In effect, this implies that the minimum feature size for thick oxide is  $3\lambda$ .

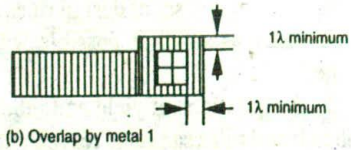
The simplicity of the lambda-based rules makes this approach to design an appropriate one for the novice chip designer and also, perhaps, for those applications in which we are not trying to achieve the absolute minimum area and the absolute maximum performance. Because lambda-based rules try 'to be all things to all people', they do suffer from least common denominator effects and from the upward rounding of all process line dimension parameters into integer values of lambda.



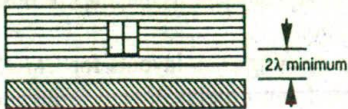
1. Aspects related to vias (double metal processes)



(a) Separation via to via



(b) Overlap by metal 1

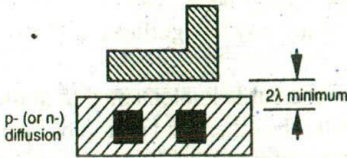


(c) Separation via to polysilicon

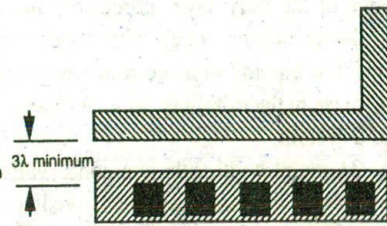


(d) Separation via to thinox

2. Polysilicon wires separation from cuts

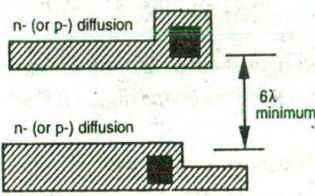


(a) Short polysilicon run



(b) Long polysilicon runs

3. Diffusion wires separation from cuts



Separations between different active areas



4. Increase in polysilicon overlap to reduce metal migration effect

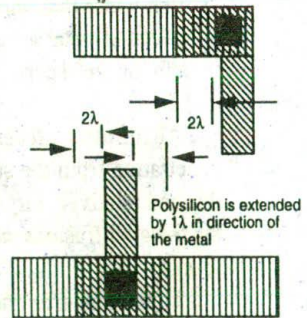


Figure 3-12 Further aspects of  $\lambda$ -based design rules for contacts, including some factors contributing to higher yield/reliability

The performance of any fabrication line in this respect clearly comes down to a matter of tolerances and definitions in terms of microns (or some other suitable unit of length). Thus, expanded sets of rules often referred to as micron-based rules are available to the more experienced designer to allow for the use of the full capability of any process. Also, many processes offer additional layers, which again adds to the possibilities presented to the designer.

In order to properly represent these important aspects, the next section introduces Orbit Semiconductor's 2  $\mu\text{m}$  feature size double metal, double poly., n-well CMOS rules which also offer a BiCMOS capability.

### 3.5 2 $\mu\text{m}$ double metal, double poly. CMOS/BiCMOS rules\*

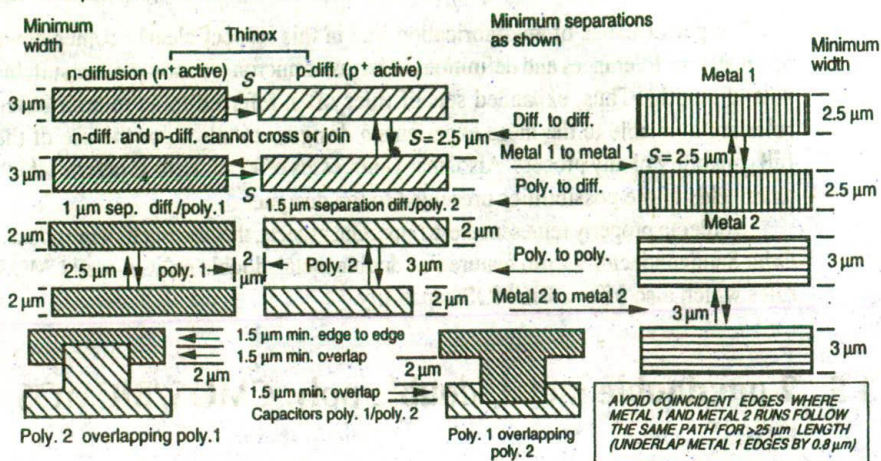
In order to accommodate the additional features present in this technology, it is necessary to extend the range of color and monochrome encodings previously used for double metal p-well CMOS. The encoding used is compatible with that already described, but as far as color assignments are concerned the following extension/additions are made: n-well — brown (same as p-well); Poly. 1 — red; Poly. 2 — orange; nDiff. (n-active) — green; pDiff. (p-active) — yellow (a green outline to the yellow may be used to show pDiff. clearly in color stick diagrams). Hatching, which is compatible with monochrome encoding, may also be added to color mask encoding, to distinguish underlying layers and to allow for ready copying of color diagrams on monochrome copying machines.

For BiCMOS the following are added: buried n<sup>+</sup> subcollector — pale green; p-base — pink. These extra features are set out in Figure 3-1(c) and in Color plate 1(c).

The use of color encoding is illustrated in the colorplates section of this book. The monochrome encoded rule set for the Orbit™ 2  $\mu\text{m}$  double metal double poly. BiCMOS process is given in Figures 3-13(a)-(f). The rule set is also presented in color as Color plates 3 to 6. Note the relative complexity of these rule sets. It must be further noted that an appropriate set of electrical parameters must accompany each set of design rules and the parameters for the Orbit™ 2  $\mu\text{m}$  process are included in Appendix A.

\* The rules and other details have been supplied by Orbit Semiconductors Inc. of Sunnyvale, California, through Integrated Silicon Design Pty Ltd of Adelaide, Australia. Their joint cooperation is gratefully acknowledged.





Otherwise polysilicon 2 must not be coincident with polysilicon 1

Note: Where no separation is specified, wires may overlap or cross (e.g. metal may cross any layer). For p-well CMOS, n-diff. wires can only exist inside and p-diff. wires outside the p-well. For n-well CMOS, p-diff. wires can only exist inside and n-diff. wires outside the n-well.

Figure 3-13(a) Design rules for wires (interconnects) (Orbit 2  $\mu\text{m}$  CMOS)

### 3.6 1.2 $\mu\text{m}$ double metal, single poly. CMOS rules\*

As fabrication technology improves, so the feature size reduces and a separate set of micron-based design rules must accompany each new feature size. In order to open up the possibilities presented by this text, we have included the Orbit™ 1.2  $\mu\text{m}$  rules in Appendix B together with the relevant electrical parameters.

### 3.7 Layout diagrams — a brief introduction

Mask layout diagrams may be hand-drawn on, say, 5 mm squared paper. In the case of lambda-based rules, the side of each square is taken to represent  $\lambda$  and, for micron-based rules, it will be taken to represent the least common factor associated with the rules (for example, 0.25  $\mu\text{m}$  per side for the 2  $\mu\text{m}$  process and

\* The rules and other related details have been supplied by Orbit Semiconductors Inc. of Sunnyvale, California, through Integrated Silicon Design Pty Ltd of Adelaide, Australia. Their joint cooperation is gratefully acknowledged.

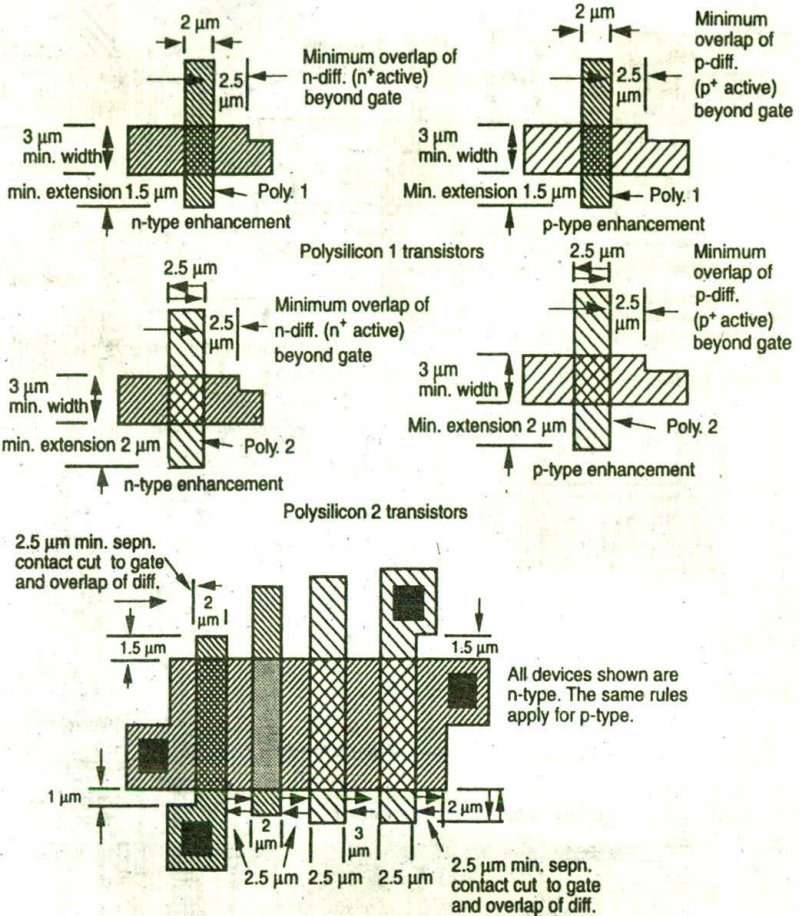


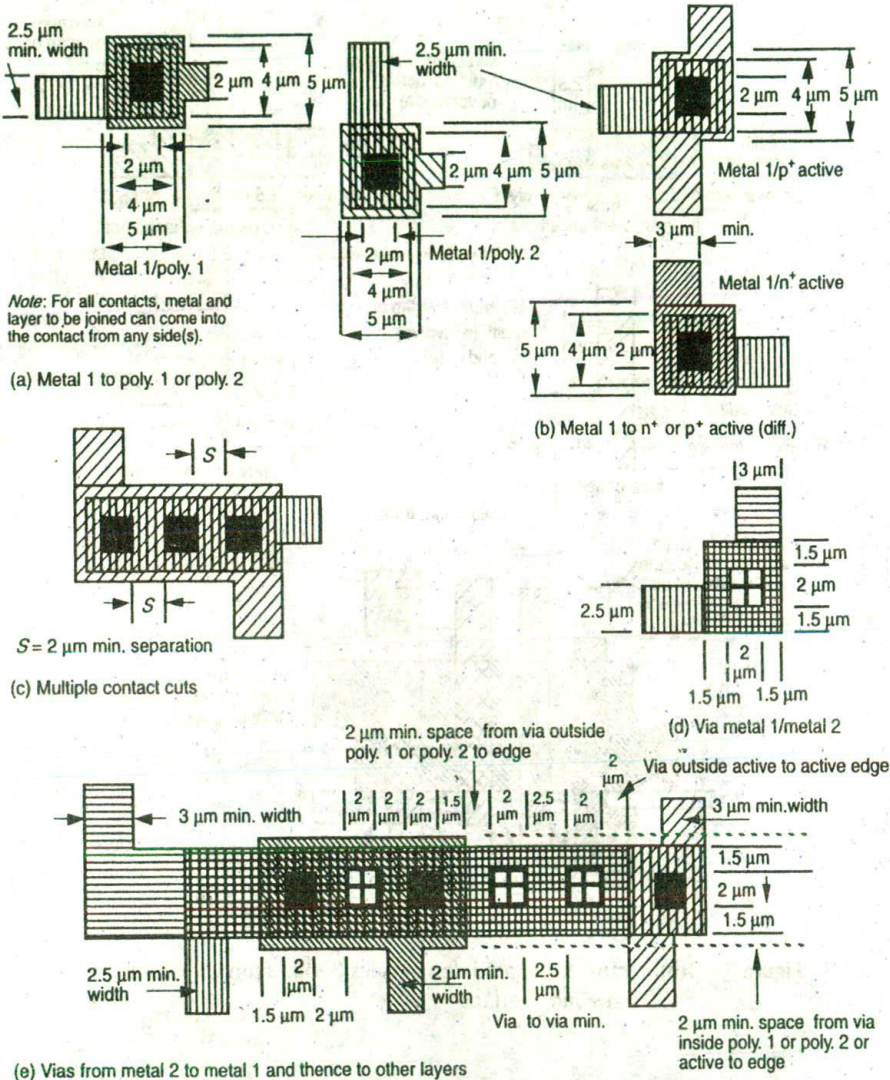
Figure 3-13(b) Transistor related design rules (Orbit 2  $\mu\text{m}$  CMOS) minimum sizes and overlaps

0.2  $\mu\text{m}$  per side for the 1.2  $\mu\text{m}$  Orbit™ process layout). Most CAD VLSI tools also offer convenient facilities for mask level design.

The introductory layout diagrams which follow in Figures 3-14 to 3-17 inclusive have been included to illustrate the use of the lambda-based rule set and many more examples will appear later in the text.

The use of butting contacts has not been illustrated here as the reader is to be discouraged from using a facility which is not widely available now, but example layouts appear elsewhere for the sake of continuity with earlier designs and previous editions of this book.





Note: For all contacts, metal and layer to be joined can come into the contact from any side(s).

Note: The vias must not be placed over contacts

Figure 3-13(c) Rules for contacts and vias (Orbit 2 μm CMOS)

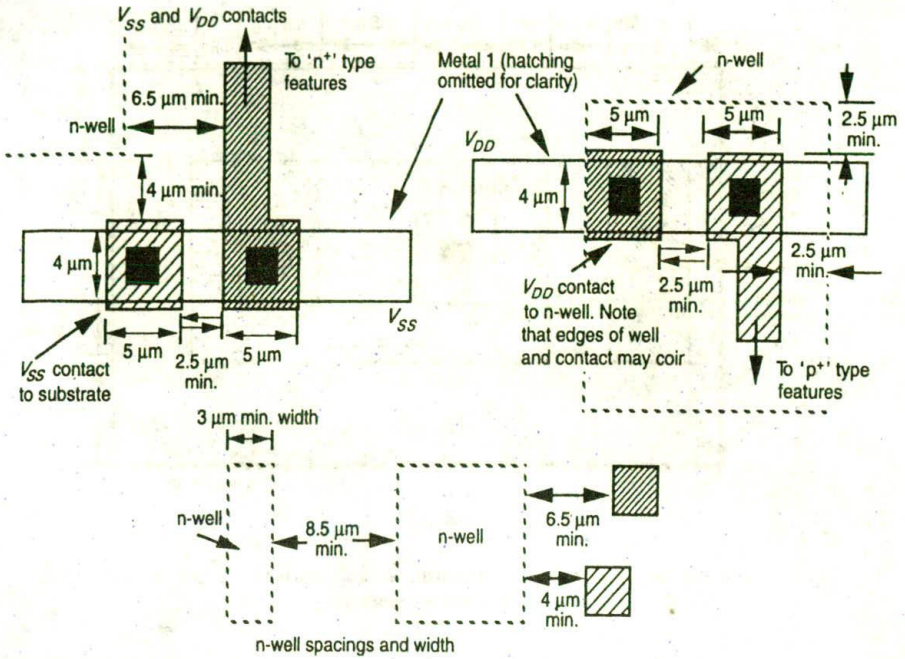
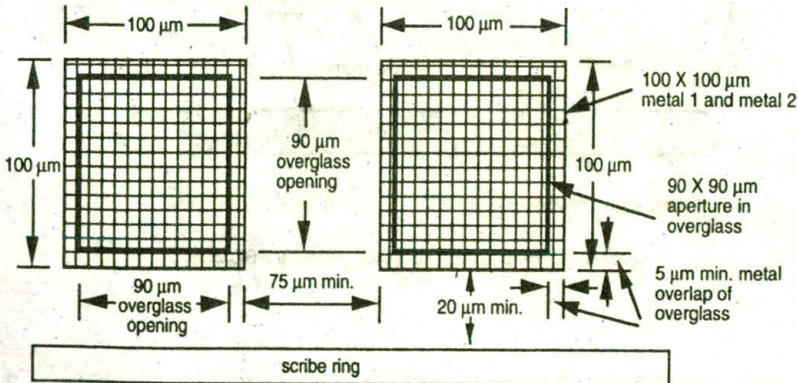


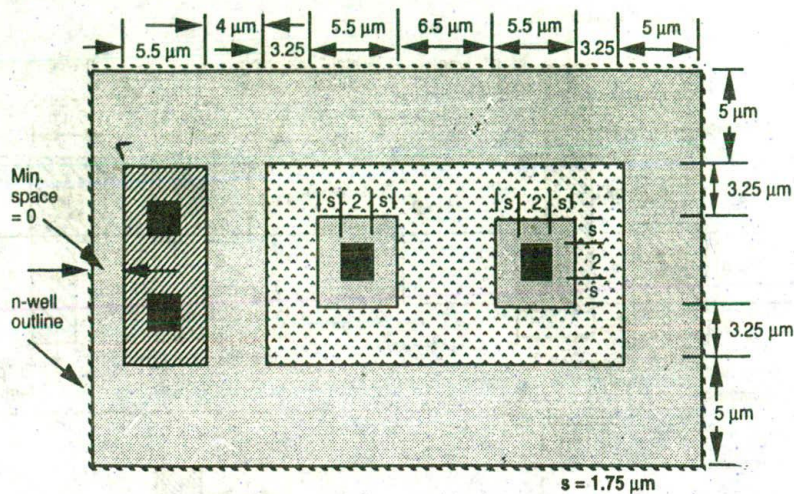
Figure 3-13(d) Rules for n-well and  $V_{DD}$  and  $V_{SS}$  contacts (Orbit 2µm CMOS)



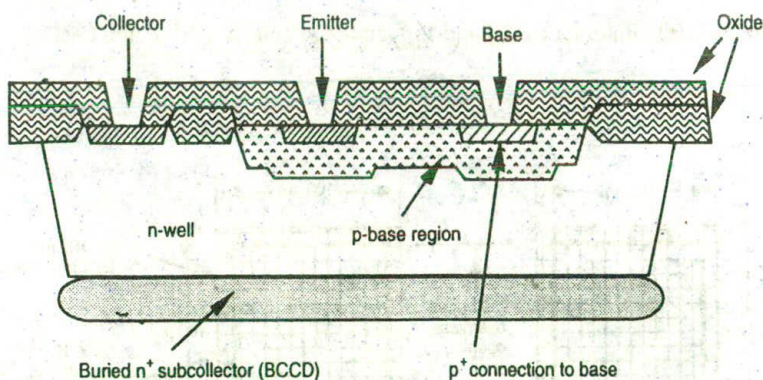
Other rules and encodings:  
 Via overlap of pad 2 µm.  
 Pad to active separation 20 µm min.  
 Color encoding for overglass mask . . . Gray

Figure 3-13(e) Rules for pad and overglass geometry (Orbit 2µm CMOS)





Note: For clarity, the layers have not been drawn transparent but BCCD underlies the entire area and the p-base underlies all within its boundary.



Cross-section through npn transistor (Orbit 2  $\mu\text{m}$  BiCMOS)

Figure 3-13(f) Special rules for BiCMOS transistors (Orbit 2  $\mu\text{m}$  CMOS)

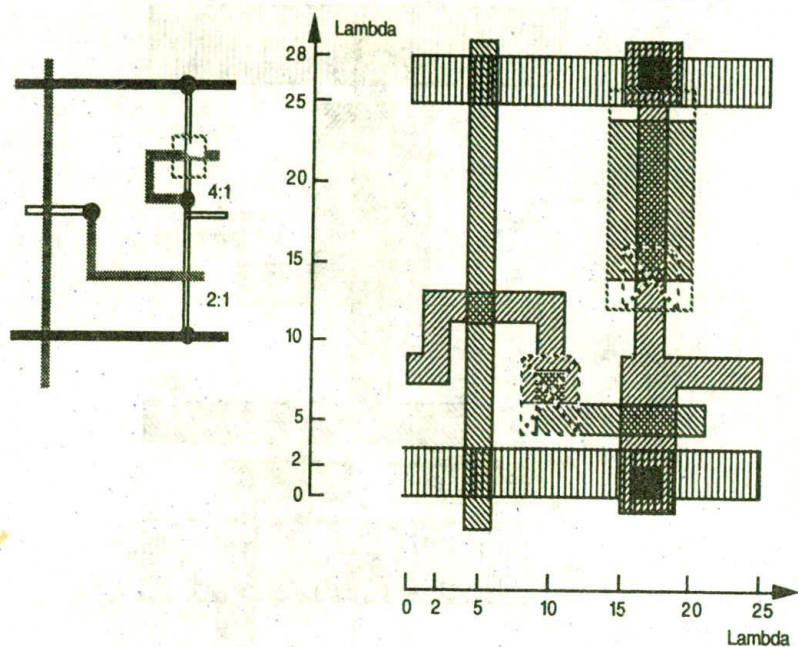


Figure 3-14 Stick diagram and layout for nMOS shift register cell

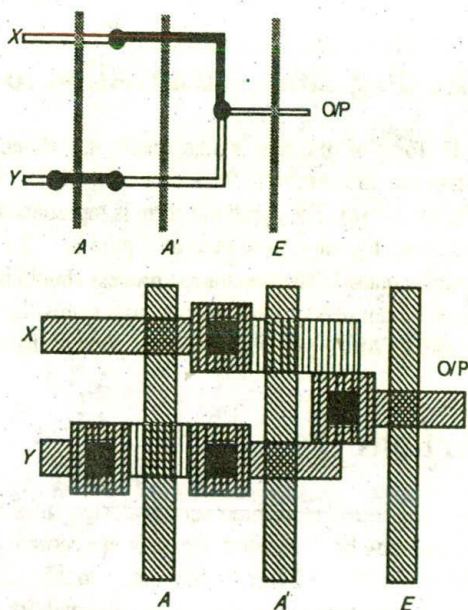


Figure 3-15 Two-way selector with enable



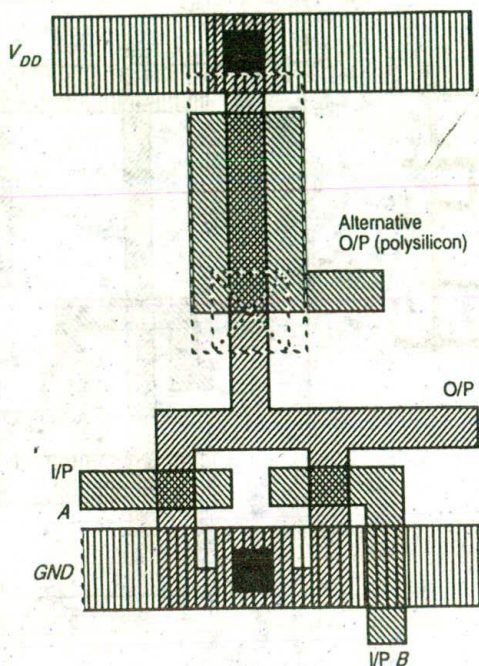


Figure 3-16 Two I/P nMOS Nor gate

### 3.8 Symbolic diagrams — translation to mask form

The symbolic form of diagram is also readily translated to mask layout form. Take, for example, the symbolic form of a 1-bit CMOS shift register cell given earlier in Figure 3-5(e). The symbolic form is reproduced in Figure 3-17(a) and the resultant mask layout is presented as Figure 3-17(b). This is also presented in color as Color plate 7. The translation process should be self-evident from the figures. Further examples of mask layout from symbolic diagram form follow in Chapter 5 (for BiCMOS gates) and in other parts of the text.

### 3.9 Observations

This chapter has introduced three sets of design rules with which nMOS and CMOS designs may be fabricated. Designs incorporating BiCMOS technology are covered by the 'Orbit' 2  $\mu\text{m}$  double metal, double poly, n-well process rules. We are now in a position to use the design rules and, for simplicity, most design examples will use the lambda-based rules. As the budding designer becomes

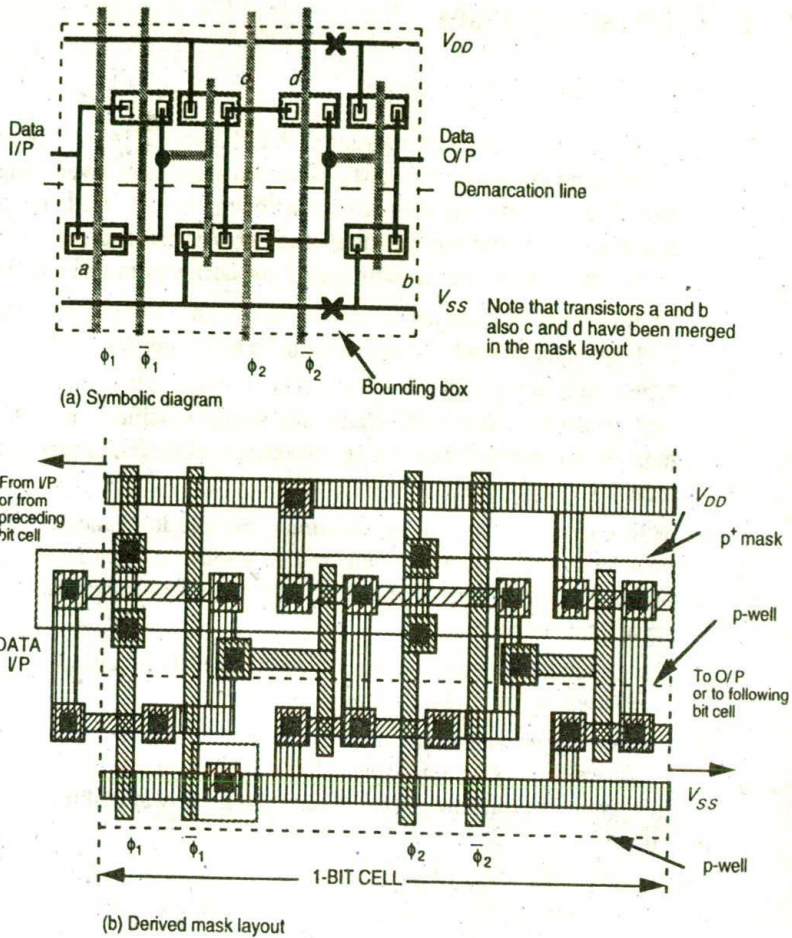


Figure 3-17 A 1-bit CMOS shift register cell

more proficient, designs are readily completed using one or other of the 'Orbit' rule sets.

Before we begin any design work, however, a further chapter is necessary to establish, explain and evaluate other key circuit parameters.



## 3.10 Tutorial exercises

*Note:* Use colors to represent layers.

1. First draw the *circuit diagrams* for each of Figures 3-14 to 3-16 and then, after closing this book, draw a stick diagram and a mask layout diagram for each. These efforts may then be compared with those in the book, although note that lack of conformity in detail may not mean that a layout, for example, is incorrect. Check your layouts against the design rules given in the text.
2. Draw the stick diagram and a mask layout for an 8:1 nMOS inverter circuit. Both the input and output points should be on the polysilicon layer.
3. With regard to Figure 3-15, what will be the state of the output (O/P) when control line E is at 0 volts? Could any simple modification to this circuit improve its operation? If so, set out a modified stick diagram and corresponding mask layout.
4. With regard to Figure 3-14, determine suitable left-hand and right-hand boundary lines for this leaf-cell, so that a series of such leaf-cells can be butted directly together side by side without violating any design rules, yet occupying minimum area.
5. Can you reduce the area occupied by the leaf-cell of Figure 3-14? Draw an alternative layout to illustrate your contention.
6. Figure 3-18 presents a simple CMOS layout. Study the layout, and from it produce a circuit diagram. Explain the nature and purpose of the circuit. Using this layout, explain how you could construct a four-way multiplexer (selector) circuit.

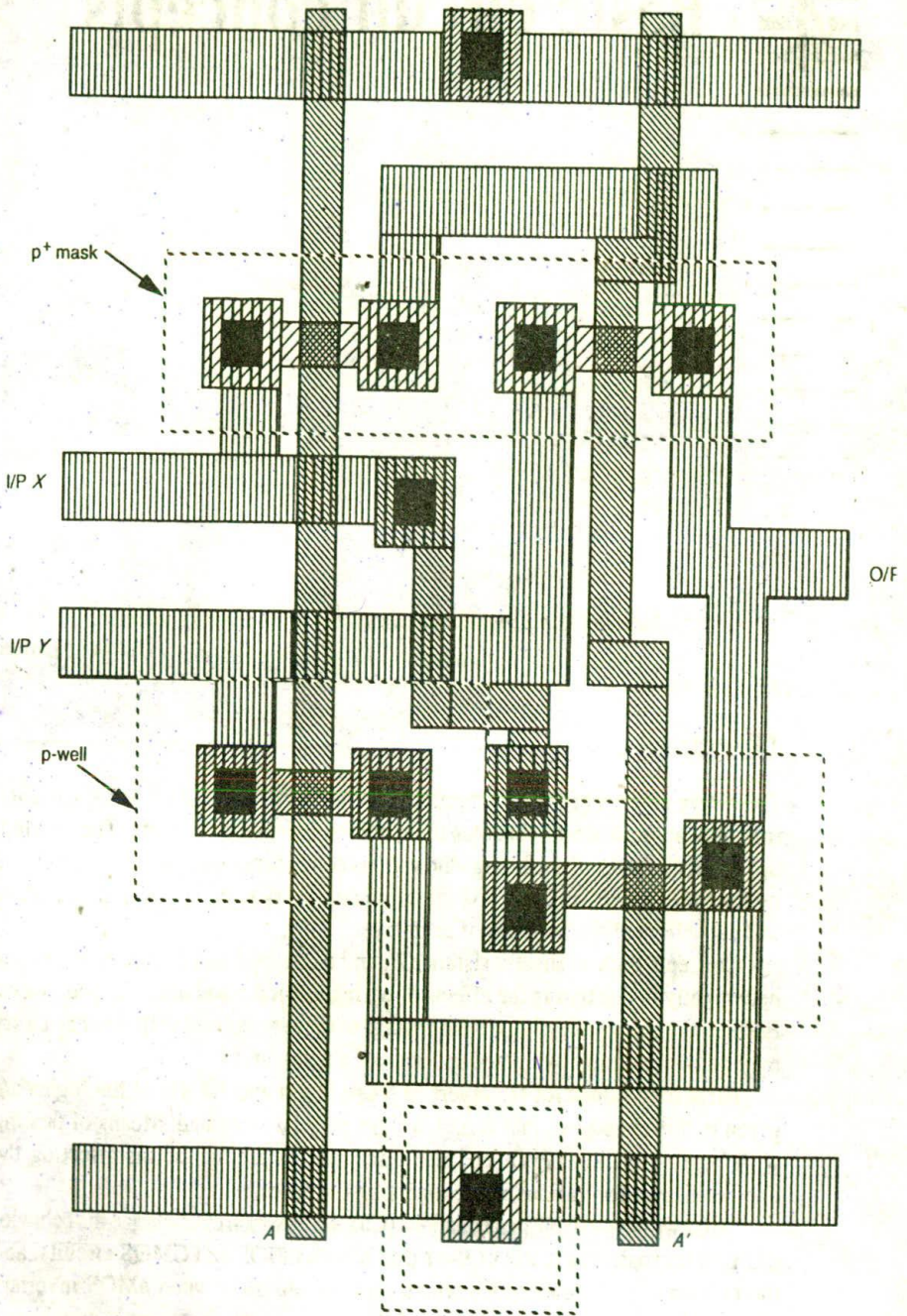


Figure 3-18 CMOS layout example



# 4

## Basic circuit concepts

*Education is a progressive discovery of our own ignorance.*

Will Durant

### Objectives

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  $\square C_g$  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  $\tau$ .

Parameter values for the layers in  $5 \mu\text{m}$ ,  $2 \mu\text{m}$  and  $1.2 \mu\text{m}$  technologies are given in this chapter so that actual designs may be evaluated. Means of dealing with larger capacitive loads are also discussed, as are the factors affecting the choice of layer for various interconnection purposes.

So far we have established equations (Chapter 2) which characterize the behavior of MOS transistors, aspects of their use in both nMOS and CMOS circuits, and the pull-up to pull-down ratios which must be observed when nMOS inverters and pass transistors are interconnected. However, as yet we have not considered the actual resistance and capacitance values associated with transistors, nor have

we considered circuit wiring and parasitics. In order to simplify the treatment of such components, there are basic circuit concepts which will now be introduced, and for particular MOS processes we can set out approximate circuit parameters which greatly ease the design process in allowing straightforward calculations. In order to take advantage of BiCMOS circuitry, we must also examine some basic properties of bipolar transistors.

## 4.1 Sheet resistance $R_s$

Consider a uniform slab of conducting material of resistivity  $\rho$ , of width  $W$ , thickness  $t$ , and length between faces  $L$ . The arrangement is shown in Figure 4-1.

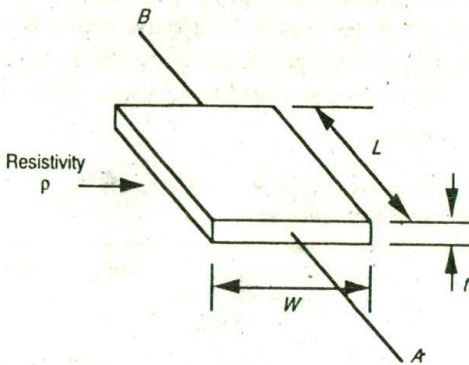


Figure 4-1 Sheet resistance model

With reference to Figure 4-1, consider the resistance  $R_{AB}$  between two opposite faces.

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

where

$$A = \text{cross-section area}$$

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

$$R_s = \text{ohm per square or sheet resistance}$$

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  $\mu\text{m}$  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  $\mu\text{m}^*$ , and Orbit 2  $\mu\text{m}^*$  and 1.2  $\mu\text{m}^*$  technologies

| Layer                   | $R_s$ , ohm per square |                     |                         |
|-------------------------|------------------------|---------------------|-------------------------|
|                         | 5 $\mu\text{m}$        | Orbit               | Orbit 1.2 $\mu\text{m}$ |
| 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    | $10^4$ †               | $2 \times 10^4$ †   | $2 \times 10^4$ †       |
| p-transistor channel    | $2.5 \times 10^4$ †    | $4.5 \times 10^4$ † | $4.5 \times 10^4$ †     |

Note: In some processes a silicide layer is used in place of polysilicon.

\* 5 micron ( $\mu\text{m}$ ) technology implies minimum line width (and feature size) of 5  $\mu\text{m}$  and in consequence  $\lambda = 2.5 \mu\text{m}$ . Similarly, 2  $\mu\text{m}$  and 1.2  $\mu\text{m}$  technologies have feature sizes of 2  $\mu\text{m}$  and 1.2  $\mu\text{m}$  respectively.

\*\* The figures given are for n-diffusion regions. The values for p-diffusion are 2.5 times these values.

† These values are approximations only. Resistances may be calculated from a knowledge of  $V_{ds}$  and the expressions for  $r_{ds}$  given earlier.

## 4.2 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.

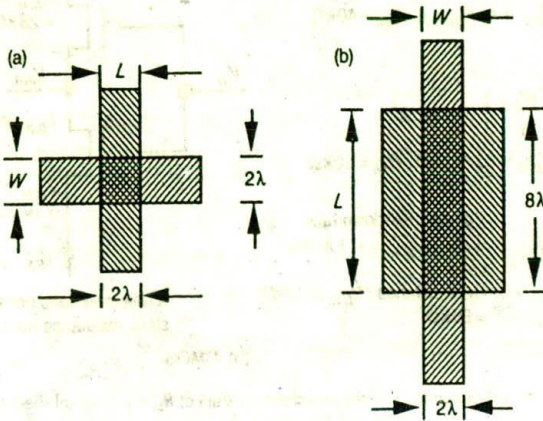


Figure 4-2 Resistance calculation for transistor channels

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)

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

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

\*From Table 4-1.



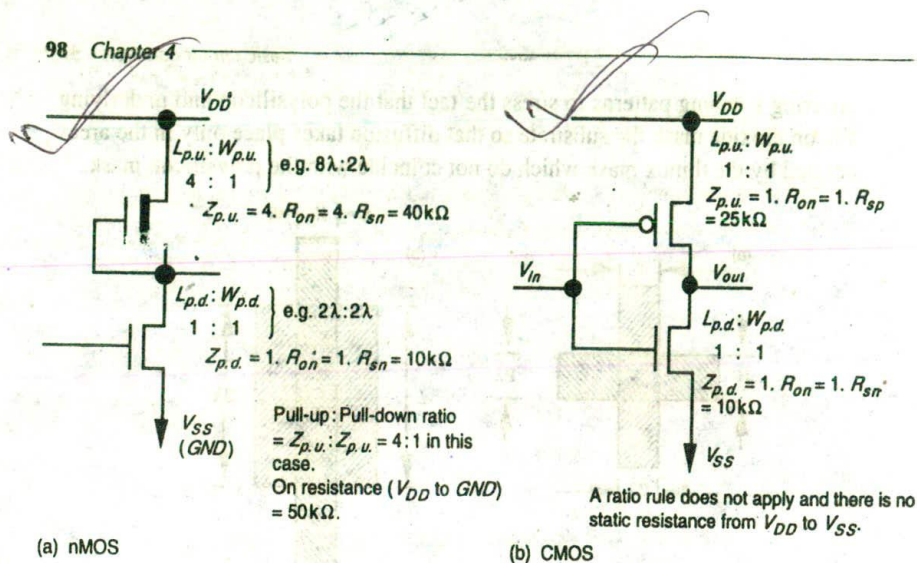


Figure 4-3 Inverter resistance calculation

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.

### 4.2.1 Silicides

As the line width becomes smaller, the sheet resistance contribution to RC delay increases. With the currently available polysilicon sheet resistance ranging from 15 to 100 ohm it is apparent that some of the advantages of scaling could be offset by the interconnect resistance at the gate level. Therefore the low sheet resistances of refractory silicides (2–4 ohm), which are formed by depositing metal on polysilicon and then sintering, have been investigated as an interconnecting medium.

Deposition of the metal or metal/silicon alloy prior to sintering may be done in any one of several ways:

- sputtering or evaporation;
- co-sputtering metal and silicon in the desired ratio from two independent targets;
- co-evaporation from the elements.

Although the properties of silicides make them attractive alternatives to polysilicon, there are extra processing steps which offset this advantage.

### 4.3 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:

$$C = \frac{\epsilon_0 \epsilon_{ins} A}{D} \text{ farads}$$

where

$D$  = thickness of silicon dioxide

$A$  = area of plates

(and it is assumed that  $\epsilon_0$ ,  $A$ , and  $D$  are in compatible units, for example,  $\epsilon_0$  in farads/cm,  $A$  in  $\text{cm}^2$ ,  $D$  in cm).

$\epsilon_{ins}$  = relative permittivity of  $\text{SiO}_2 \doteq 4.0$

$\epsilon_0 = 8.85 \times 10^{-14}$  F/cm (permittivity of free space)

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

$$C \left( \frac{\text{pF}}{\mu\text{m}^2} \right) = \frac{\epsilon_0 \epsilon_{ins}}{D} \frac{\text{F}}{\text{cm}^2} \times \frac{10^{12} \text{pF}}{\text{F}} \times \frac{\text{cm}^2}{10^8 \mu\text{m}^2}$$

( $D$  in cm,  $\epsilon_0$  in farads/cm)

Typical values of area capacitance are set out in Table 4-2 for 5  $\mu\text{m}$  technology and for Orbit 2  $\mu\text{m}$  and 1.2  $\mu\text{m}$  technologies.

Table 4-2 Typical area capacitance values for MOS circuits

| Capacitance               | Value in $\text{pF} \times 10^{-4}/\mu\text{m}^2$ (Relative values in brackets) |                 |                   |
|---------------------------|---|-----------------|-------------------|
|                           | 5 $\mu\text{m}$   | 2 $\mu\text{m}$ | 1.2 $\mu\text{m}$ |
| 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).



## 4.4 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 as 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. (This concept, originated by VTI (USA), has been adapted here.)

$\square C_g$  may be evaluated for any MOS process. For example, for 5  $\mu\text{m}$  MOS circuits:

Area/standard square =  $5 \mu\text{m} \times 5 \mu\text{m} = 25 \mu\text{m}^2$  (= area of minimum size transistor)

Capacitance value (from Table 4-2) =  $4 \times 10^{-4} \text{ pF}/\mu\text{m}^2$

Thus, standard value  $\square C_g = 25 \mu\text{m}^2 \times 4 \times 10^{-4} \text{ pF}/\mu\text{m}^2 = .01 \text{ pF}$

or, for 2  $\mu\text{m}$  MOS circuits (Orbit):

Area/standard square =  $2 \mu\text{m} \times 2 \mu\text{m} = 4 \mu\text{m}^2$

Gate capacitance value (from Table 4-2) =  $8 \times 10^{-4} \text{ pF}/\mu\text{m}^2$

Thus, standard value  $\square C_g = 4 \mu\text{m}^2 \times 8 \times 10^{-4} \text{ pF}/\mu\text{m}^2 = .0032 \text{ pF}$

and, for 1.2  $\mu\text{m}$  MOS circuits (Orbit):

Area/standard square =  $1.2 \mu\text{m} \times 1.2 \mu\text{m} = 1.44 \mu\text{m}^2$

Gate capacitance value (from Table 4-2) =  $16 \times 10^{-4} \text{ pF}/\mu\text{m}^2$

Thus, standard value  $\square C_g = 1.44 \mu\text{m}^2 \times 16 \times 10^{-4} \text{ pF}/\mu\text{m}^2 = .0023 \text{ pF}$

## 4.5 Some area capacitance calculations

The approach will be demonstrated using  $\lambda$ -based geometry. The calculation of capacitance values may now be undertaken by establishing the ratio between the area of interest and the area of standard (feature size square) gate ( $2\lambda \times 2\lambda$  for  $\lambda$ -based rules) and multiplying this ratio by the appropriate relative  $C$  value from Table 4-2. The product will give the required capacitance in  $\square C_g$  units.

Consider the area defined in Figure 4-4. First, we must calculate the area relative to that of a standard gate.

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

Now:

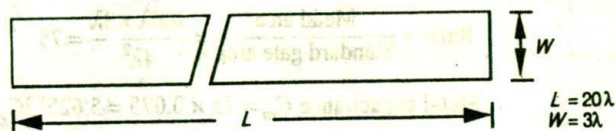


Figure 4-4 Simple area for capacitance calculation

1. Consider the area in metal 1.

$$\begin{aligned}\text{Capacitance to substrate} &= \text{relative area} \times \text{relative } C \text{ value} \\ &= 15 \times 0.075 \square C_g \\ &= 1.125 \square C_g\end{aligned}$$

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.

$$\begin{aligned}\text{Capacitance to substrate} &= 15 \times 0.1 \square C_g \\ &= 1.5 \square C_g\end{aligned}$$

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

$$\begin{aligned}\text{Capacitance to substrate} &= 15 \times 0.25 \square C_g \\ &= 3.75 \square C_g^*\end{aligned}$$

Calculations of area capacitance values associated with structures occupying more than one layer, as in Figure 4-5, are equally straightforward.

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

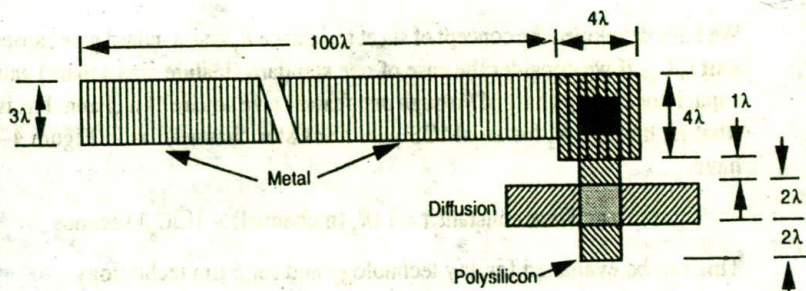


Figure 4-5 Capacitance calculation (multilayer)

\* Note the relatively high capacitance values of the diffusion layer even though peripheral capacitance (see Table 4-3 in section 4.10.3) has not been allowed for. This may increase total diffusion capacitance to considerably more than the area capacitance calculated here.



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

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

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,

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

Therefore

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

In all cases absolute values are readily evaluated by substitution of the actual value for  $\square C_g$  as given in section 4.4.

It is not unusual to find metal paths of uniform  $4\lambda$  width but when taking this approach in design it must be borne in mind that, compared with  $3\lambda$  width paths, the capacitance will be increased by one-third.

For example, if the metal width is increased to  $4\lambda$  in Figure 4-5, the capacitance  $C_m$  is increased to  $7.5 \square C_g$  and the capacitance of the complete structure will increase to about  $9 \square C_g$ .

## 4.6 The delay unit $\tau$

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:

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

This can be evaluated for any technology and for  $5 \mu\text{m}$  technology,

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

and for  $2 \mu\text{m}$  (Orbit) technology,

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

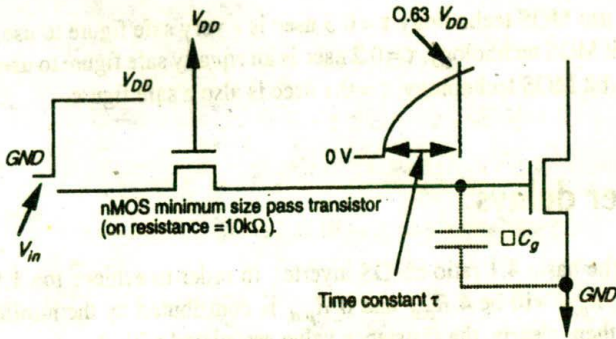


Figure 4-6 Model for derivation of  $\tau$

and for 1.2  $\mu\text{m}$  (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  $\tau$  is often increased by a factor of two or three so that for 5  $\mu\text{m}$  circuit

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

Note that  $\tau$  thus obtained is not much different from transit time  $\tau_{sd}$  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  $\tau$  in Figure 4-6, so that an appropriate value for  $V_{ds}$  is the average value = 3 volts. For 5  $\mu\text{m}$  technology, then,

$$\begin{aligned} \tau_{sd} &= \frac{25 \mu\text{m}^2 \text{ V sec}}{650 \text{ cm}^2 \text{ 3V}} \times \frac{10^9 \text{ nsec cm}^2}{10^8 \mu\text{m}^2} \\ &= 0.13 \text{ nsec} \end{aligned}$$

This is very close to the theoretical time constant  $\tau$  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  $\tau$ ) interchangeably and 'stray' capacitances are usually allowed for by doubling (or more) the theoretical values calculated.

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



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

## 4.7 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  $5\tau$ .

However, a single 4:1 inverter exhibits undesirable asymmetric delays since the delay in turning on is, for example,  $\tau$ , while the corresponding delay in turning off is  $4\tau$ . 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-

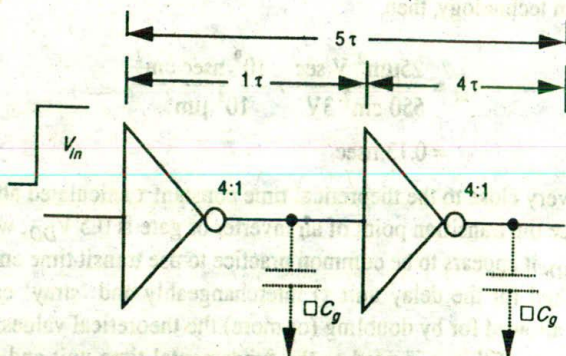


Figure 4-7 nMOS inverter pair delay

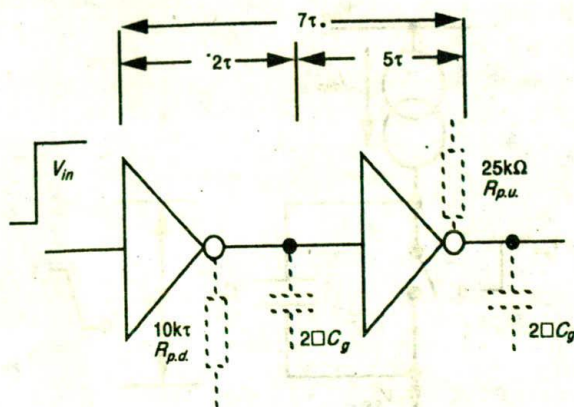


Figure 4-8 Minimum size CMOS inverter pair delay

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 ( $= 2C_g$ ) is double that of the comparable nMOS inverter since the input to a CMOS inverter is connected to *both* transistor gates. Note also the allowance made for the differing channel resistances.

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

#### 4.7.1 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  $\tau_r$  or fall-time  $\tau_f$  can be estimated from the following simple analysis.

##### 4.7.1.1 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

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

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



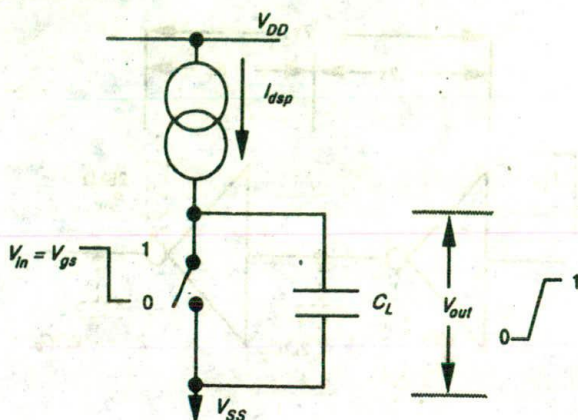


Figure 4-9 Rise-time model

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

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  $V_{out} \doteq +V_{DD}$ , so that

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

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

#### 4.7.1.2 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.

Making similar assumptions we may write for fall-time:

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

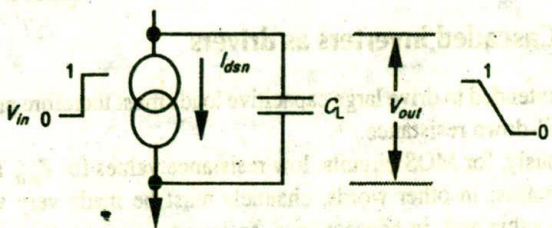


Figure 4-10 Full-time model

#### 4.7.1.3 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  $\beta_n \doteq 2.5 \beta_p$ , 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  $W_p = 2.5 W_n$  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.  $\tau_r$  and  $\tau_f$  are proportional to  $1/V_{DD}$ ;
2.  $\tau_r$  and  $\tau_f$  are proportional to  $C_L$ ;
3.  $\tau_r = 2.5 \tau_f$  for equal n- and p-transistor geometries.

## 4.8 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.



### 4.8.1 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).

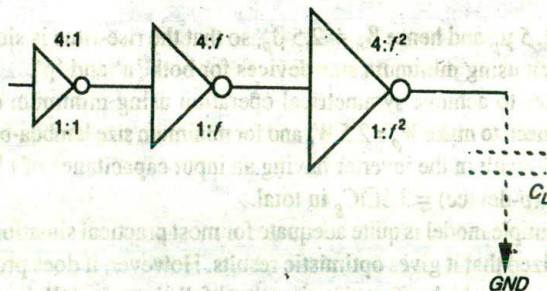


Figure 4-11 Driving large capacitive loads

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

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

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

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

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  $t_d$

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

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

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

or

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

\* Usually, a value of  $f = 3$  is used since the curve showing delay versus  $f$  is quite flat around the minimum.



### 4.8.2 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.

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.

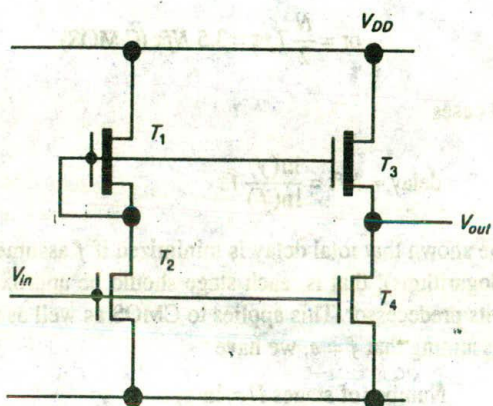


Figure 4-12 Inverting type nMOS super buffer

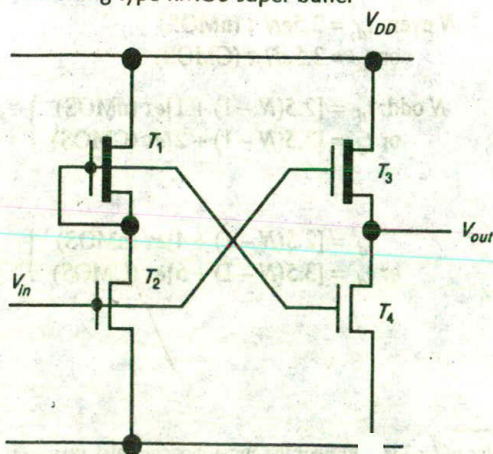


Figure 4-13 Non-inverting type nMOS super buffer

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  $I_{ds} \propto V_{gs}$ , 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  $\mu\text{m}$  technology are capable of driving loads of 2 pF with 5 nsec rise-time.

Other nMOS arrangements such as those based on the native transistor, and known as native super buffers, may be used, but such processes are not readily available to the designer and are mentioned here only briefly.

### 4.8.3 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 (Chapter 2) 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  $\mu_n$  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.

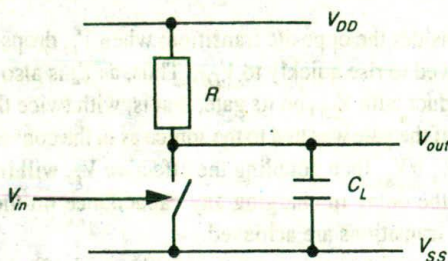
The switching performance of a transistor driving a capacitive load may be visualized initially from the simple model given in Figure 4-14.

It may be shown that the time  $\Delta t$  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.





Note: The time necessary to change the output voltage by an amount that is equal to the input change is given by

$$\Delta t = C_L/g_m$$

where

$g_m$  = device transconductance.

Figure 4-14 Driving ability of bipolar transistor

Clearly, since the bipolar transistor has a relatively high transconductance, the value of  $\Delta t$  is small.

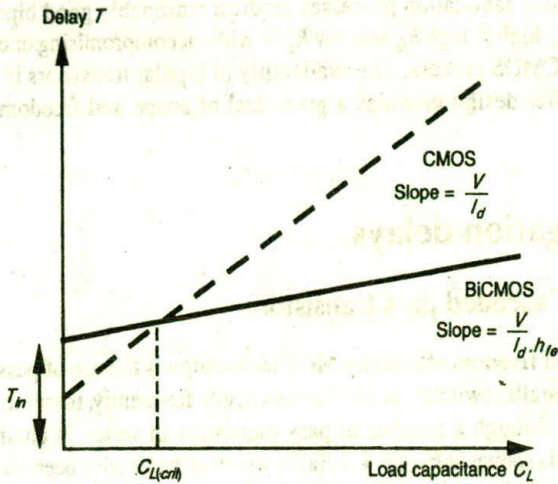
A more exacting appraisal of the bipolar transistor delay reveals that it comprises two main components:

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.

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.



- 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

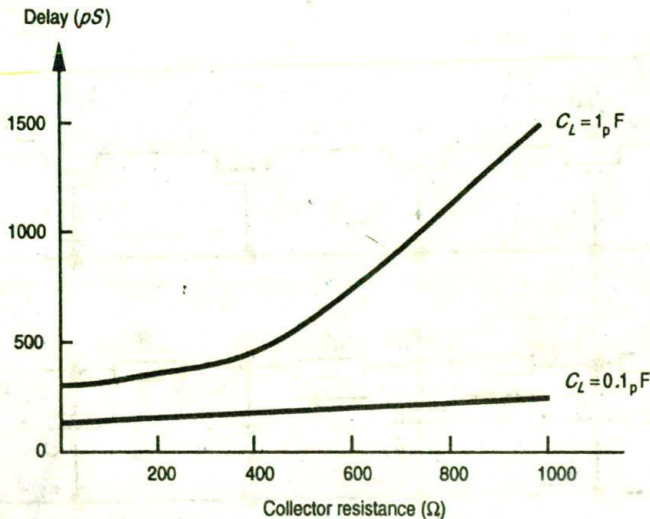


Figure 4-16 Gate delay as a function of collector resistance



BiCMOS fabrication processes produce reasonably good bipolar transistors — high  $g_m$ , high  $\beta$ , high  $h_{fe}$  and low  $R_c$  — without compromising or overlaborating 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.

## 4.9 Propagation delays

### 4.9.1 Cascaded pass transistors

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

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

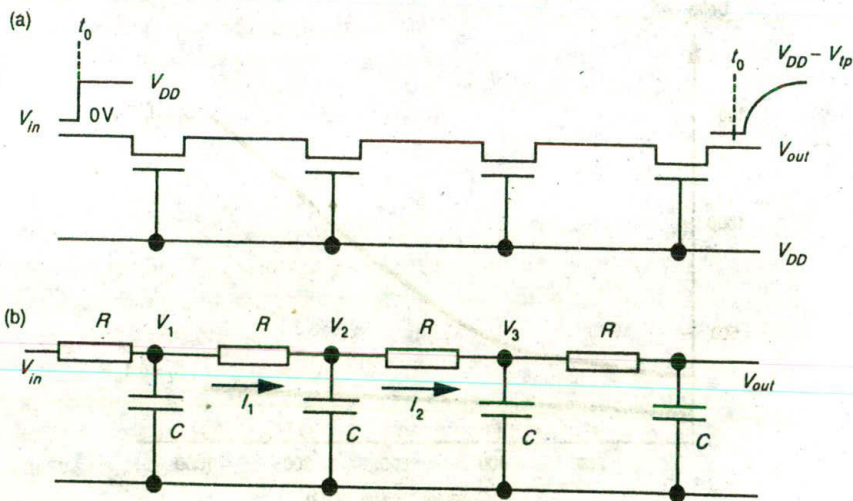


Figure 4-17 Propagation delays in pass transistor chain

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.

The propagation time  $\tau_p$  for a signal to propagate a distance  $x$  is such that

$$\tau_p \propto x^2$$

The analysis can be simplified if all  $R$ s and  $C$ s 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  $\tau_d$  for  $n$  sections is given by

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

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.

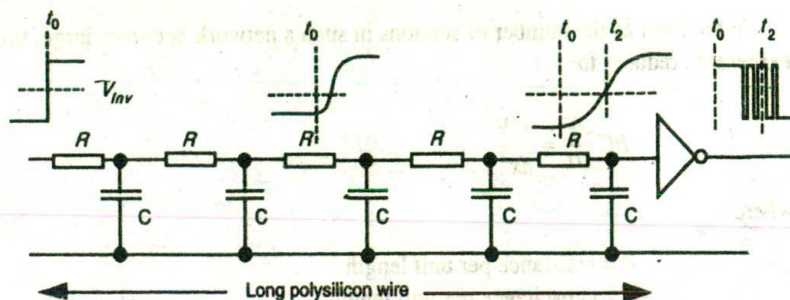
## 4.9.2 Design of long polysilicon wires

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





Note:  $V_{inv}$  = Inverter threshold

Figure 4-18 Possible effects of delays in long polysilicon wires

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.

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.

## 4.10 Wiring capacitances

In section 4.5 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.

### 4.10.1 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.

$$C_{ff} = \epsilon_{SiO_2} \epsilon_0 l \left[ \frac{\pi}{1n \left\{ 1 + \frac{2d}{t} (1 + \sqrt{1 + \frac{t}{d}}) \right\}} - \frac{t}{4d} \right]$$

where

$l$  = wire length

$t$  = thickness of wire

$d$  = wire to substrate separation

Then, total wire capacitance

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

### 4.10.2 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.

### 4.10.3 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

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

Typical values follow in Table 4-3. For further considerations on capacitive effects the reader is referred to Arpad Barna, *VHSIC — Technologies and Tradeoffs*, Wiley, 1981



Table 4-3 Typical values for diffusion capacitances

| Diffusion capacitance                             | Typical values                           |   |   |
|---|--|---|---|
|   | 5 $\mu\text{m}$                          | 2 $\mu\text{m}$                           | 1.2 $\mu\text{m}$                         |
| Area C ( $C_{\text{area}}$ )<br>(as in Table 4-2) | $1.0 \times 10^{-4}$ pF/ $\mu\text{m}^2$ | $1.75 \times 10^{-4}$ pF/ $\mu\text{m}^2$ | $3.75 \times 10^{-4}$ pF/ $\mu\text{m}^2$ |
| Periphery ( $C_{\text{periph}}$ )                 | $8.0 \times 10^{-4}$ pF/ $\mu\text{m}$   | negligible*                               | negligible*                               |

\* Assuming implanted regions of negligible depth.

## 4.11 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 ( $5 \mu\text{m}$ ) | $\mu\text{m}$ -based ( $2 \mu\text{m}$ ) | $\mu\text{m}$ -based ( $1.2 \mu\text{m}$ ) |
| Metal              | chip wide                            | chip wide                                | chip wide                                  |
| Silicide           | $2,000\lambda$                       | NA                                       | NA   |
| Polysilicon        | $200\lambda$                         | $400 \mu\text{m}$                        | $250 \mu\text{m}$                          |
| Diffusion (active) | $20\lambda^*$                        | $100 \mu\text{m}$                        | $60 \mu\text{m}$                           |

\* Taking account of peripheral and area capacitances. NA = not applicable.

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.   |

## 4.12 Observations

This chapter has completed our examination of the factors determining the characteristics and performance of MOS circuits in silicon. Useful concepts have been introduced and tables of typical parameter values have been set out to allow ready estimation of the performance of simple designs. Methods of dealing with larger capacitive loads, for example 'off chip' loads, have also been discussed.

All the basic information for carrying out and evaluating simple design work is now in place and will be put into practice following a discussion on scaling effects.



## 4.13 Tutorial exercises

1. A particular layer of MOS circuit has a resistivity  $\rho = 1 \text{ ohm cm}$ . A section of this layer is  $55 \mu\text{m}$  long and  $5 \mu\text{m}$  wide and has a thickness of  $1 \mu\text{m}$ . Calculate the resistance from one end of this section to the other (along the length). Use the concept of sheet resistance  $R_s$ . What is the value of  $R_s$ ?
2. A particular section of a layout (as in Figure 4-19) includes a  $3\lambda$  wide metal path which crosses a  $2\lambda$  wide polysilicon path at right angles. Assuming that the layers are separated by a  $0.5 \mu\text{m}$  thick layer of silicon dioxide, find the capacitance between the two layers.

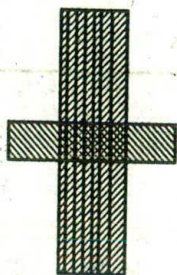


Figure 4-19 Layout detail for Question 2

The polysilicon layer in turn crosses a  $4\lambda$  wide diffusion region at right angles to form a transistor. Using the tables provided in the text, find the gate to channel capacitance. Compare it with the metal to polysilicon capacitance already calculated.

Assume  $\lambda = 2.5 \mu\text{m}$  in all cases.

3. 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  $\tau$  for the inverter geometry indicated in the figure. What are the ratios of each inverter? If strays and wiring are allowed for, it would be reasonable to increase the capacitance to ground across the output of each inverter by  $4 \square C_g$ . What is the pair delay allowing for strays? Assume a suitable value for  $\tau$  and evaluate this pair delay.
4. An off chip capacitance load of  $5\text{pF}$  is to be driven from (a) CMOS and (b) nMOS inverters. Set out suitable arrangements giving appropriate channel  $L:W$  ratios and dimensions. Calculate the number of inverter stages required, and the delay exhibited by the overall arrangement driving the  $5\text{pF}$  load.

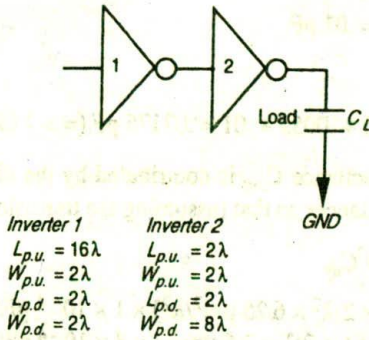


Figure 4-20 Circuit for Question 3

5. A worked example: Using the parameters given in this chapter calculate the  $C_{in}$  and  $C_{out}$  values of capacitance for the structure represented in Figure 4-21

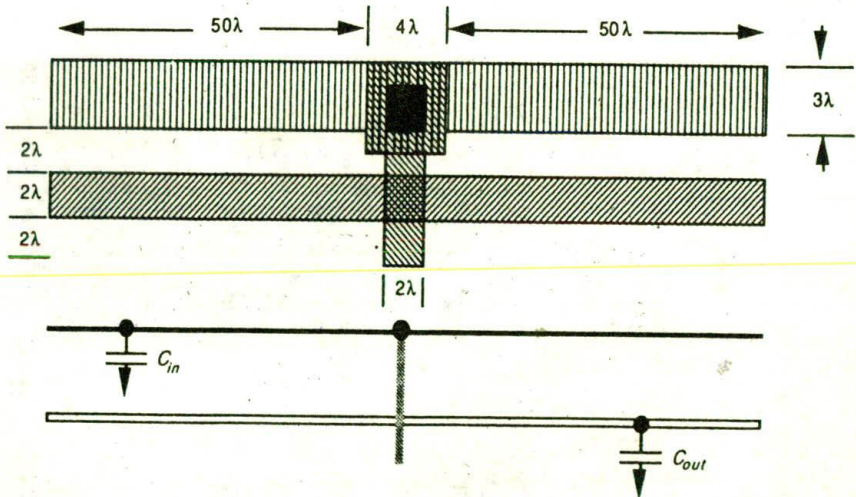


Figure 4-21 Structure for Question 5

*Solution:* The input capacitance  $C_{in}$  is made up of three components — metal bus capacitance  $C_m$ , polysilicon capacitance  $C_p$ , and the gate capacitance  $C_g$ . Thus

$$C_{in} = C_m + C_p + C_g$$

$$C_m = [2 \times (50 \times 3)\lambda^2 \times 6.25 \mu\text{m}^2/\lambda^2] \{0.3 \times 10^{-4} \text{ pF}/\mu\text{m}^2\}$$

$$= .05625 \text{ pF}$$

$$C_p = [(4 \times 4 + 2 \times 2 + 2 \times 1)\lambda^2 \times 6.25 \mu\text{m}^2/\lambda^2] \{0.4 \times 10^{-4} \text{ pF}/\mu\text{m}^2\}$$

$$= .0055 \text{ pF}$$



$$C_g = 1 \square C_g = .01 \text{ pF}$$

Thus

$$C_{in} = .05625 + .0055 + .01 = .07175 \text{ pF } (= 7 \square C_g)$$

Now, the output capacitance  $C_{out}$  is contributed by the diffusion area  $C_{da}$  and peripheral  $C_{dp}$  capacitances so that (assuming the transistor is off) we have

$$\begin{aligned} C_{out} &= C_{da} + C_{dp} \\ &= [(51 \times 2)\lambda^2 \times 6.25 \mu\text{m}^2/\lambda^2] \times 1 \times 10^{-4} \text{ pF}/\mu\text{m}^2 + \\ &\quad [2 \times (51 + 2)\lambda \times 2.5 \mu\text{m}/\lambda] \times 8 \times 10^{-4} \text{ pF}/\mu\text{m} \\ &= .06375 + .212 = .27575 \text{ pF (note significance of } C_{dp} \text{)} \end{aligned}$$