FUTURE VISION BIE

One Stop for All Study Materials

& Lab Programs

Future Vision

By K B Hemanth Raj

Scan the QR Code to Visit the Web Page

Or

Visit : https://hemanthrajhemu.github.io

Gain Access to All Study Materials according to VTU,
CSE – Computer Science Engineering,
ISE – Information Science Engineering,
ECE - Electronics and Communication Engineering
& MORE...

Join Telegram to get Instant Updates: https://bit.ly/VTU_TELEGRAM

Contact: MAIL: futurevisionbie@gmail.com

INSTAGRAM: www.instagram.com/hemanthraj_hemu/

INSTAGRAM: www.instagram.com/futurevisionbie/

WHATSAPP SHARE: https://bit.ly/FVBIESHARE

Eastern Economy Edition

Third Edition

Basic VLSI Design

Douglas A. Pucknell
Kamran Eshraghian

PHI

Chapter

3

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.

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 (thinox) silicon dioxide insulating layers. The thin oxide (thinox) mask region includes

*Orbit Semiconductor Inc., California.

n-diffusion, p-diffusion, and transistor channels. Polysilicon and thinox regions interact so that a transistor is formed where they cross one another. In some processes, there may be a second metal layer and also, in some processes, a second polysilicon layer. Layers may deliberately joined together where contacts are formed. We have also seen that the basic MOS transistor properties can be modified by the use of an implant within the thinox 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 l(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, 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.

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 l(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 l(b). Figure 3.1(c) and Color plate l(c) further extend the representations to cover a second polysilicon layer and BiCMOS technology.

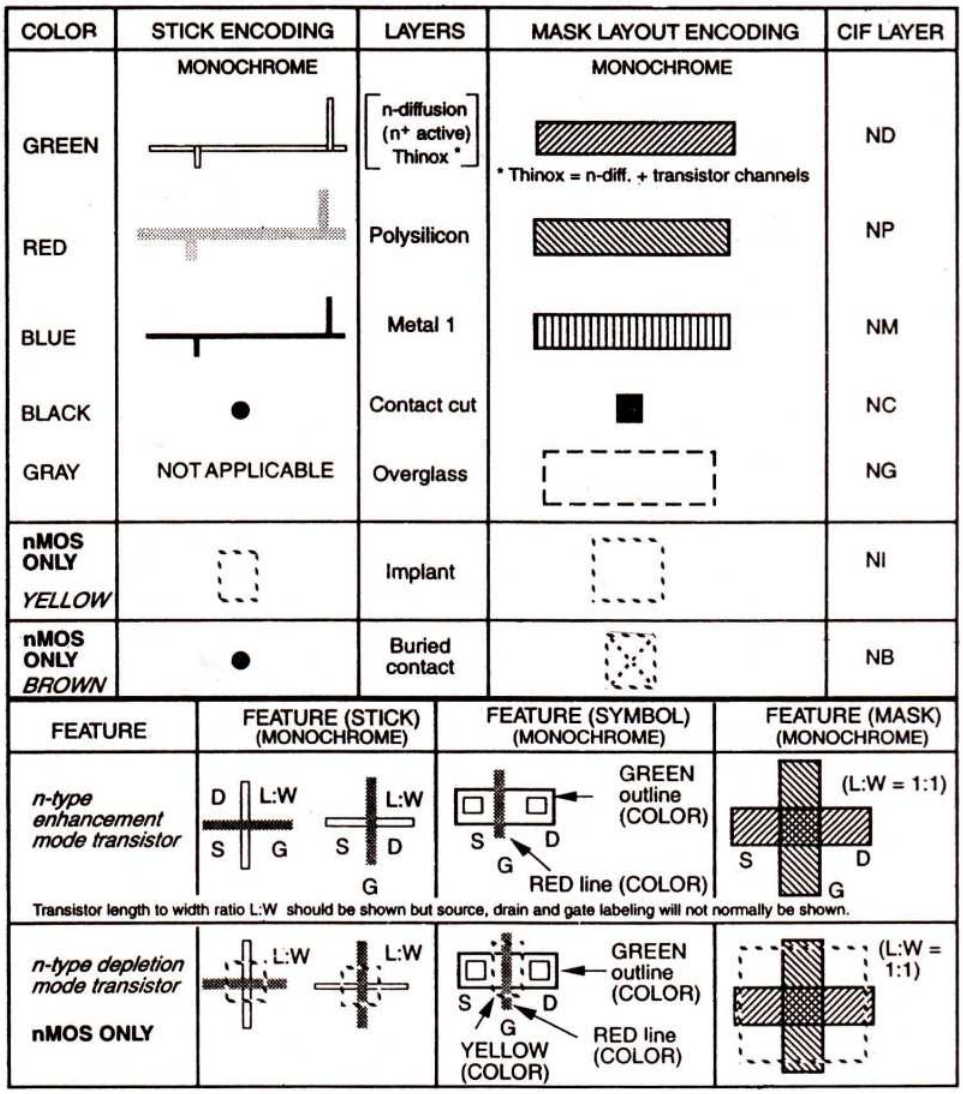
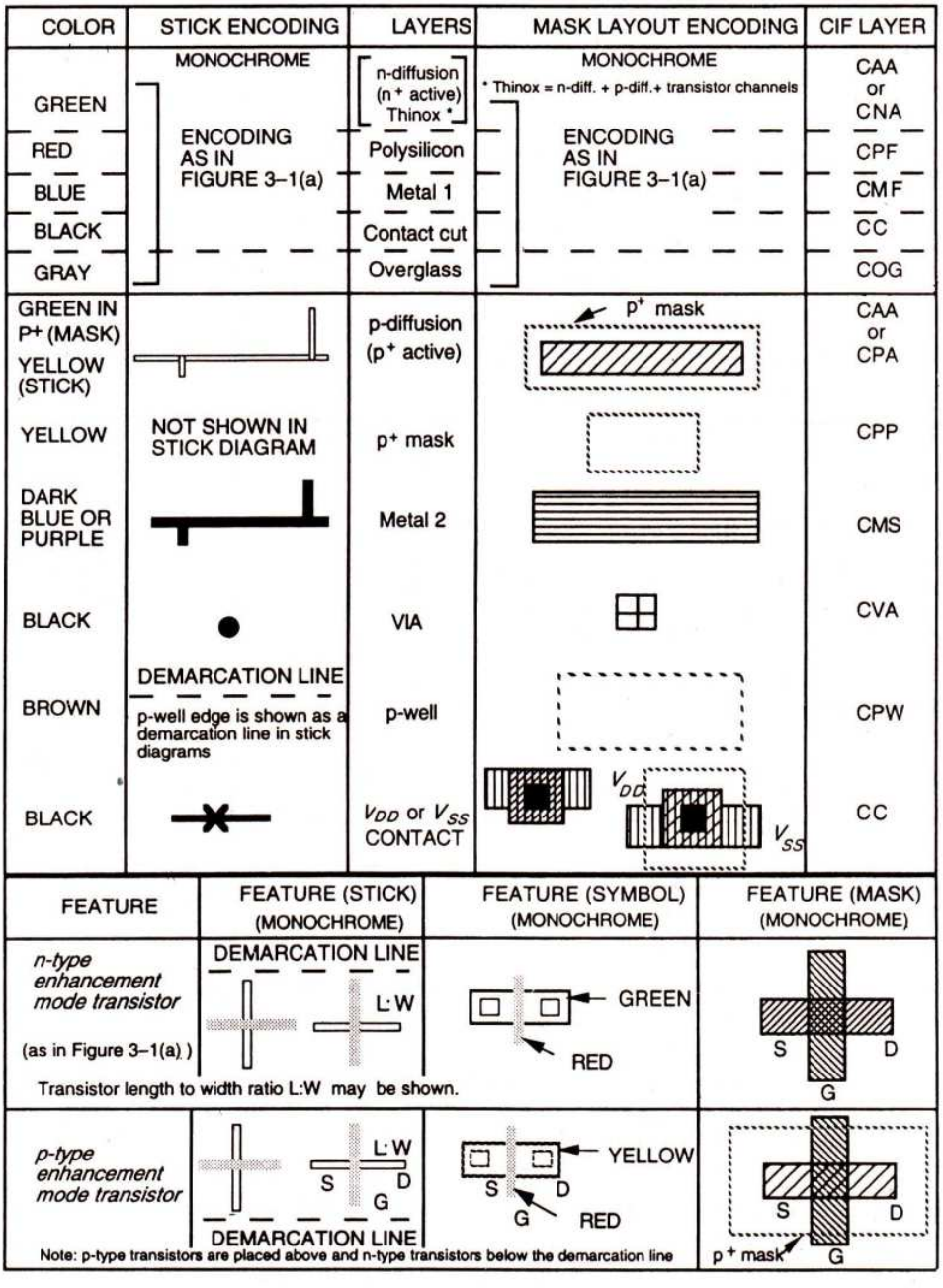


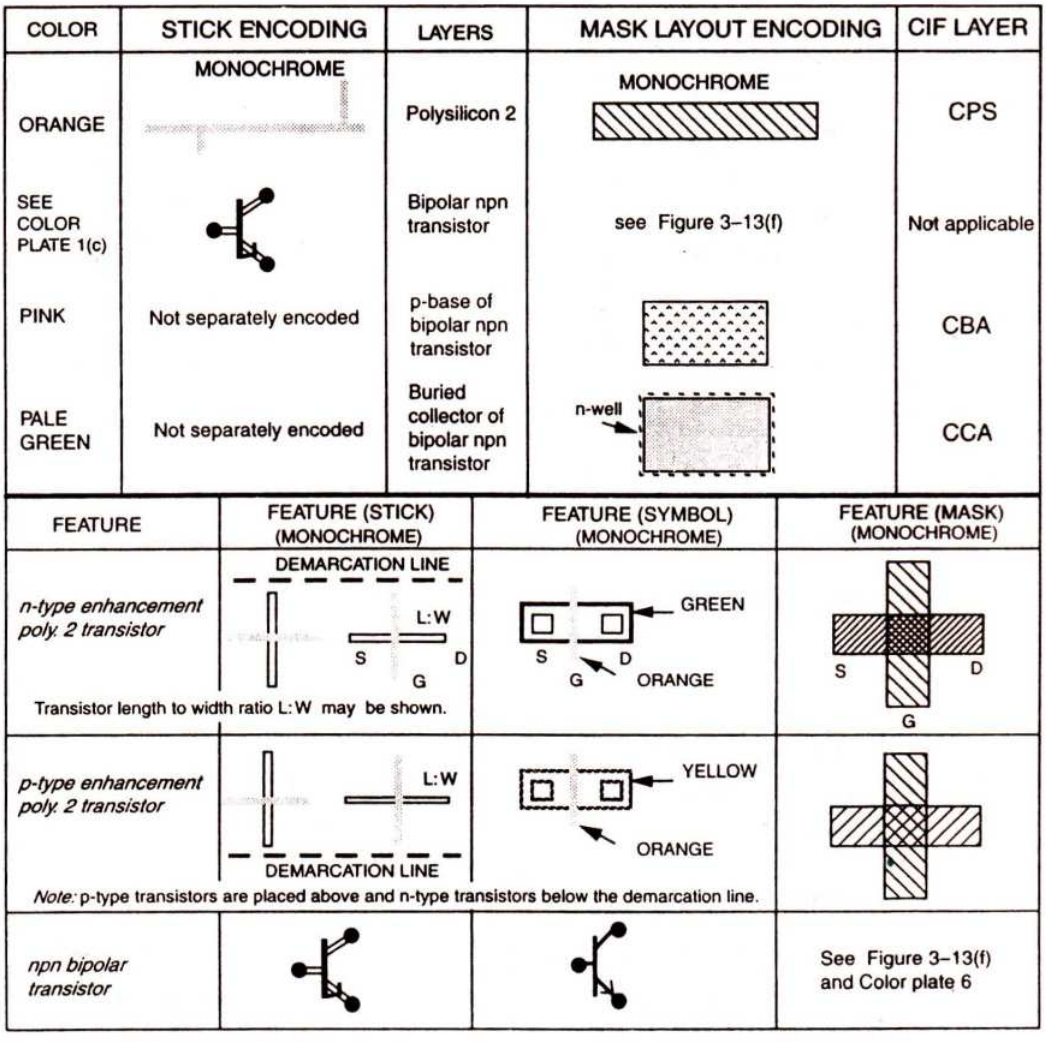
FIGURE 3.1(a) Encodings for a simple metal nMOS process (see Color plate 1(a) for nMOS color encoding details).

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 l(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 l(d), and in Color plate 7.



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).



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).

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.

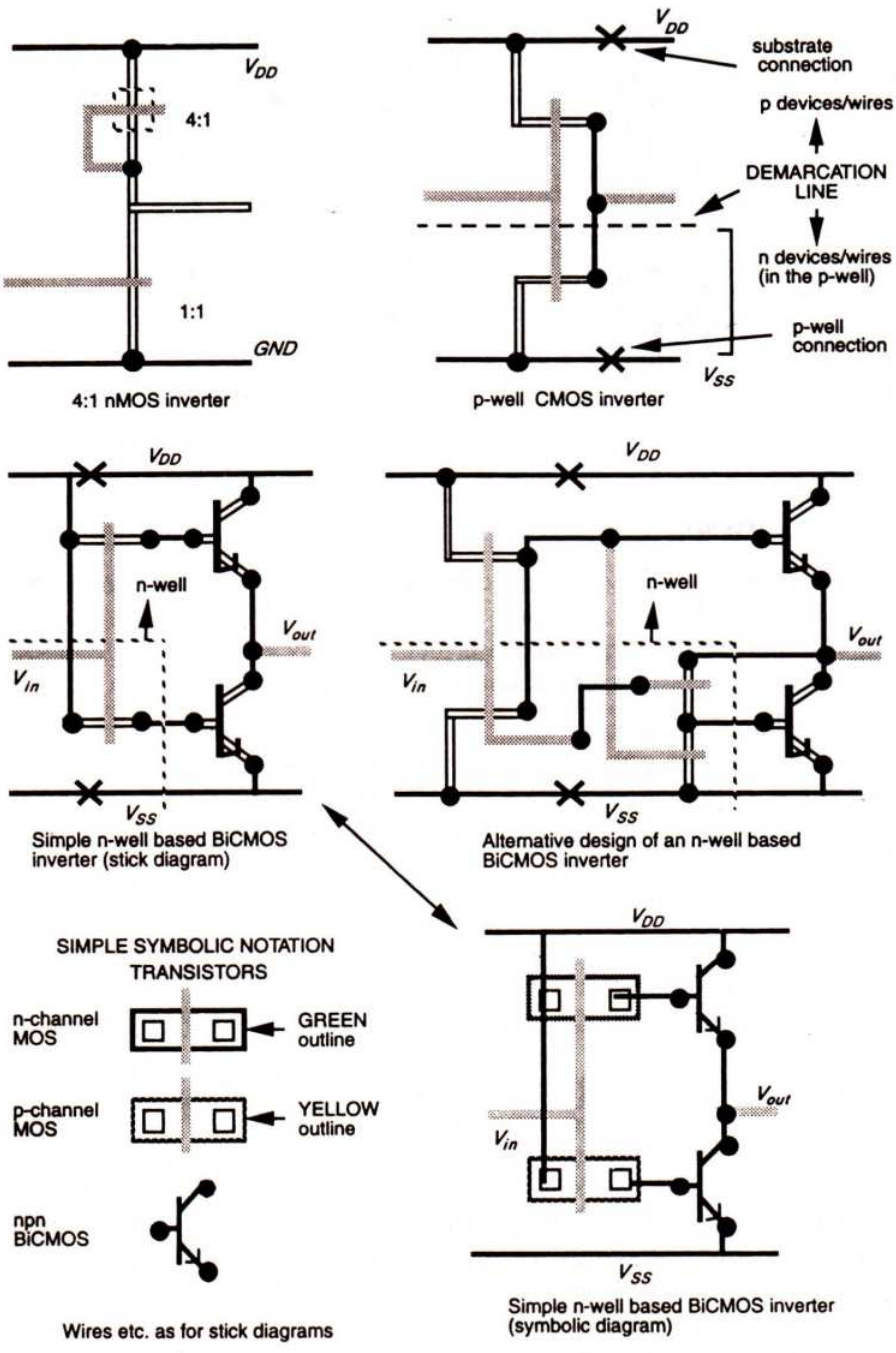


FIGURE 3.1(d) Stick diagrams and simple symbolic encoding (see also Color plate 1(d)).

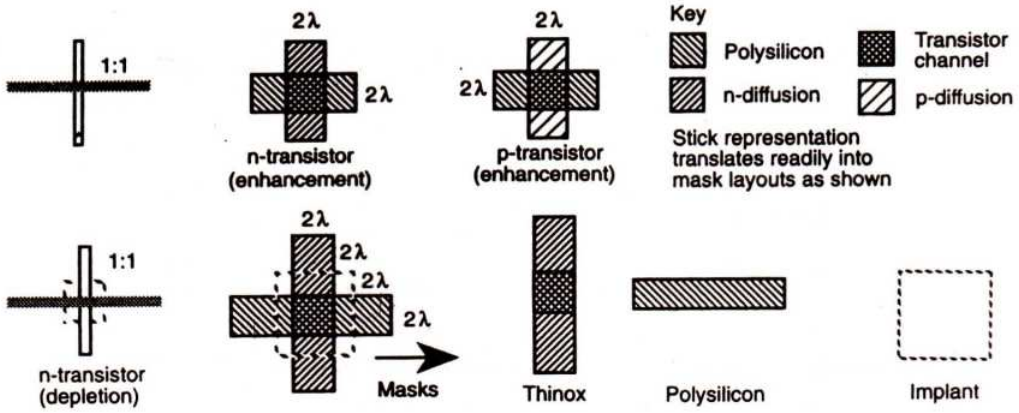


FIGURE 3.2 Stick diagrams and corresponding mask layout examples.

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 l(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

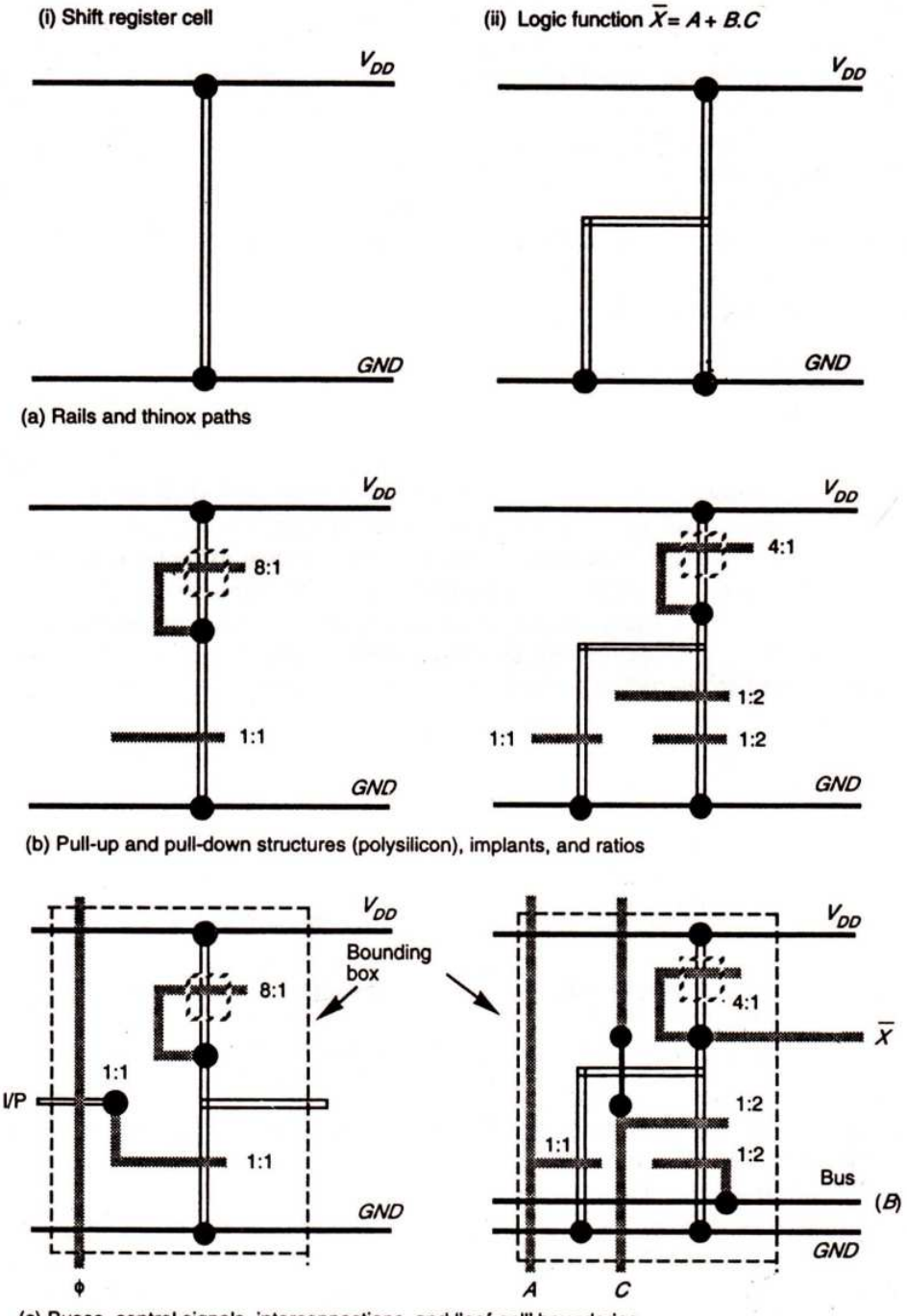


FIGURE 3.3 Examples of nMOS stick layout design style.

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.

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.

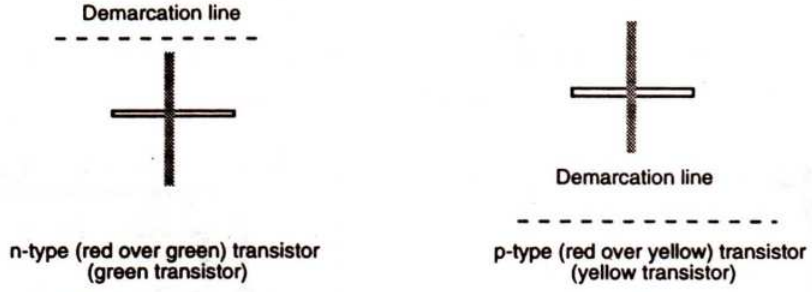
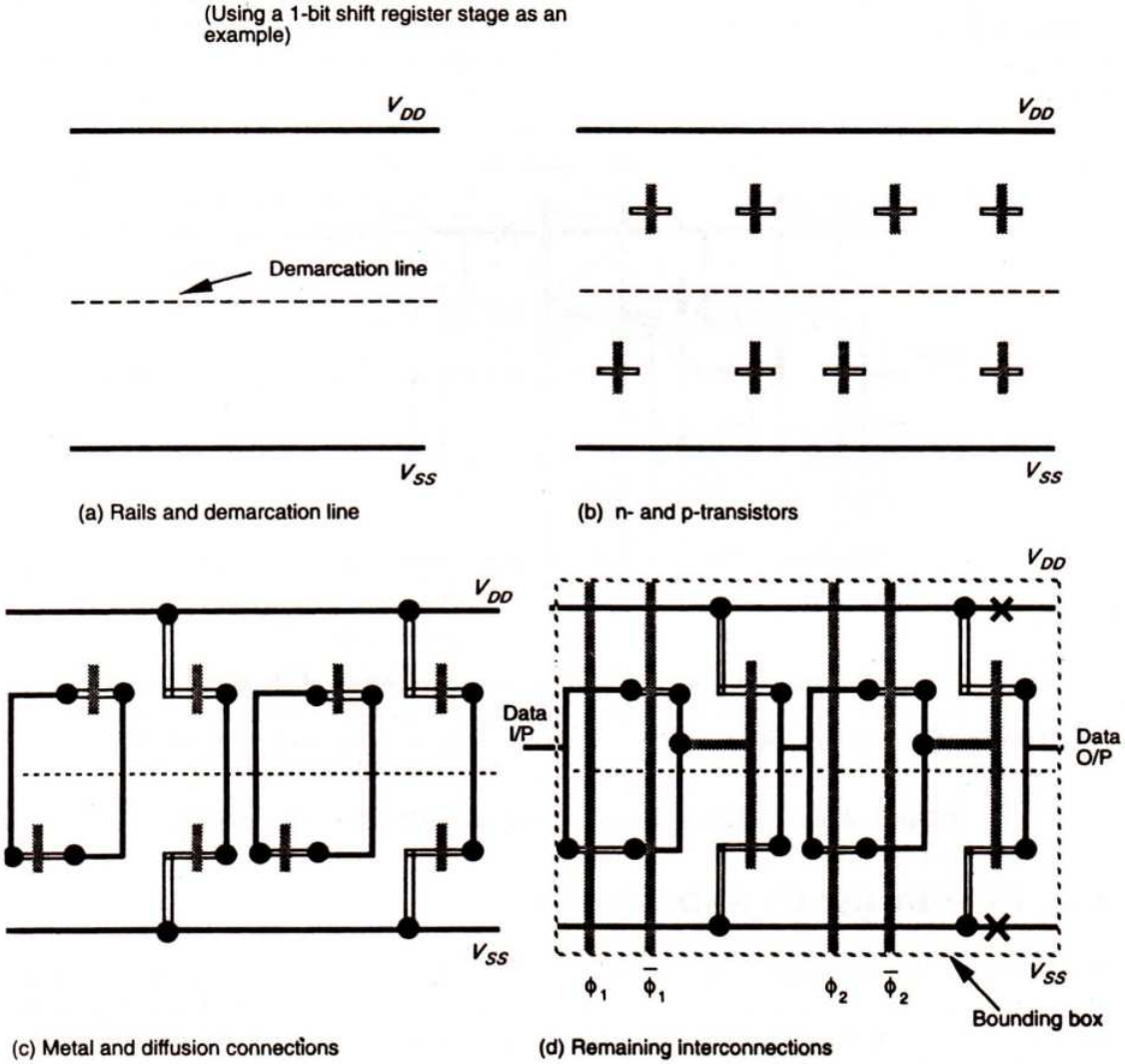


FIGURE 3.4 n-type and p-type transistors in CMOS design.

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



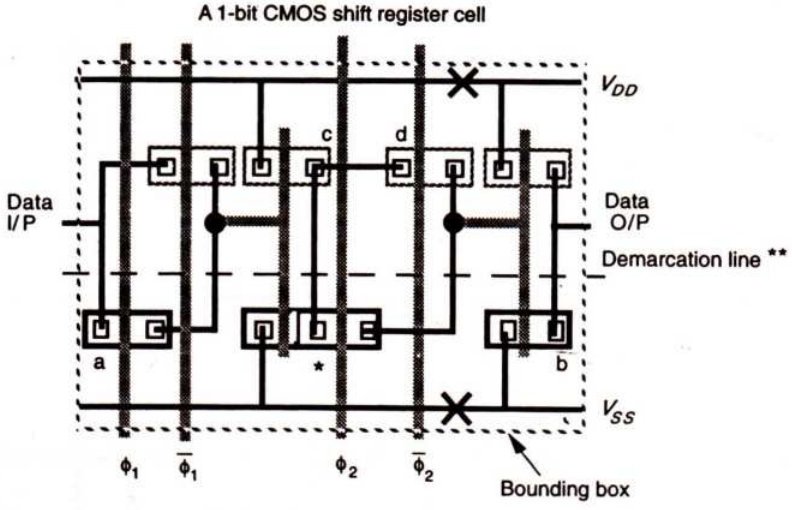
Note: The contact crosses in (d) should represent one V_{DD} contact for every four p-transistors and one V_{SS} contact for every four n-transistors.

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 (λ)-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 λ 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 λ provide a process and feature size-independent way of setting out mask dimensions to scale.

All paths in all layers will be dimensioned in λ units and subsequently λ 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 λ. For example, λ can be allocated a value of 1.0 μm so that minimum feature size on chip will be 2 μm (2λ). Design rules, also due to Mead and Conway, specify line widths, separations, and extensions in terms of λ, 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.

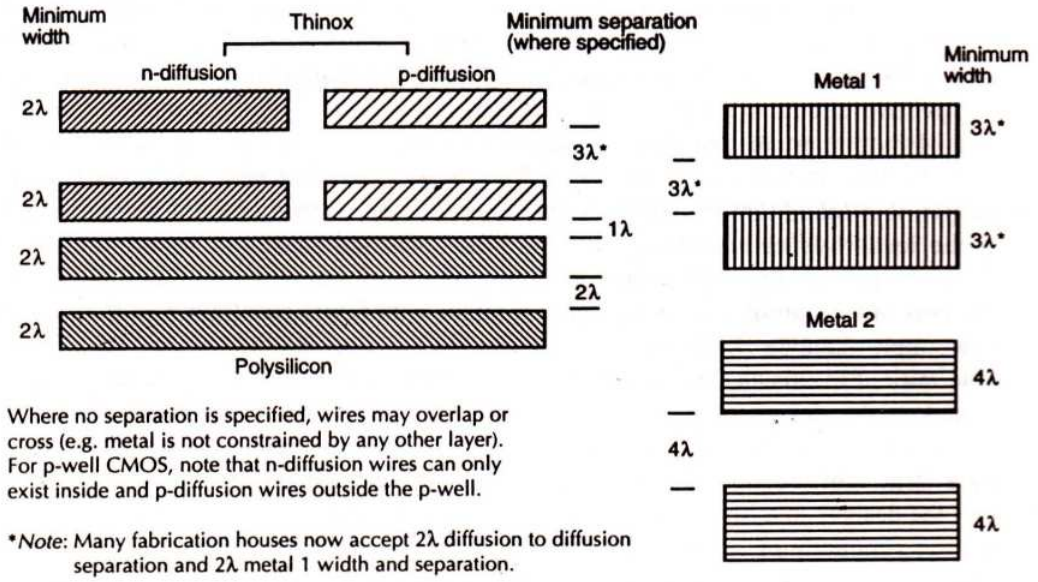


Figure 3–6 Design rules for wires (nMOS and CMOS)

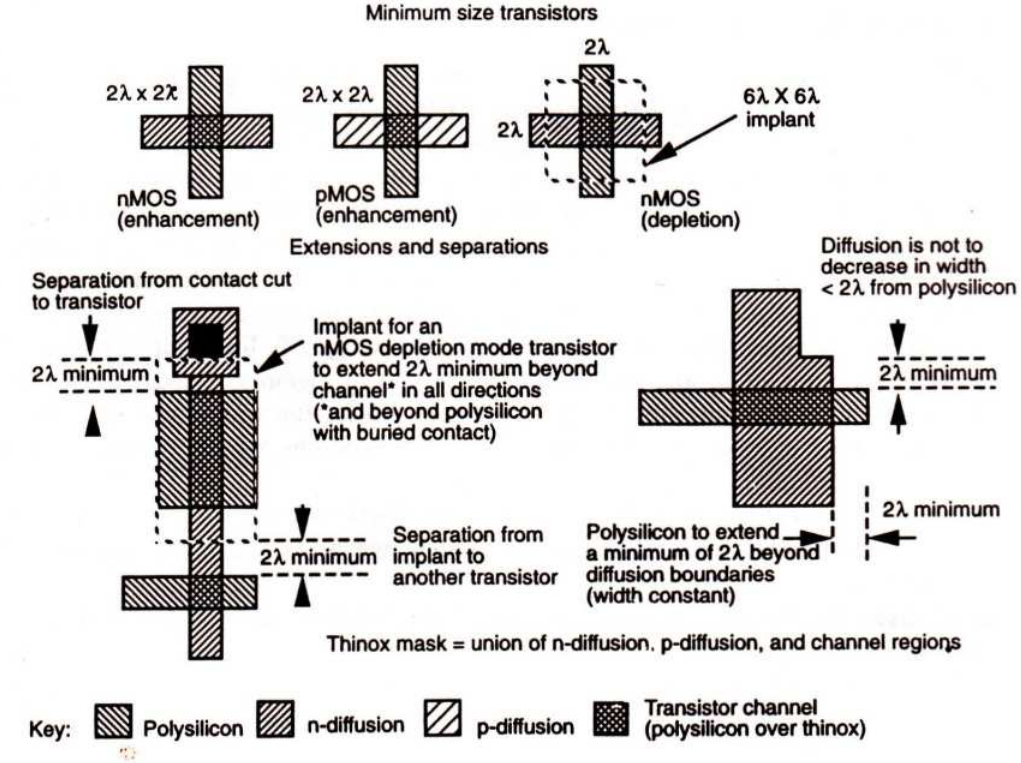


FIGURE 3.7 Transistor design rules (nMOS, pMOS and CMOS).

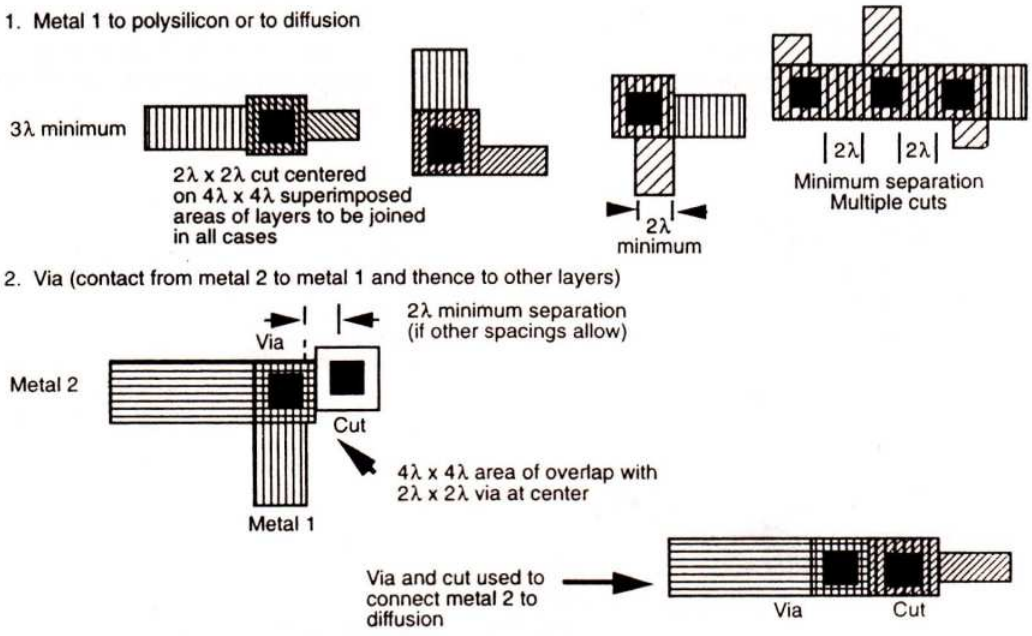


FIGURE 3.8 Contacts (nMOS and CMOS).

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 $2\lambda \times 2\lambda$ 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 $2\lambda \times 2\lambda$ 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 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.

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

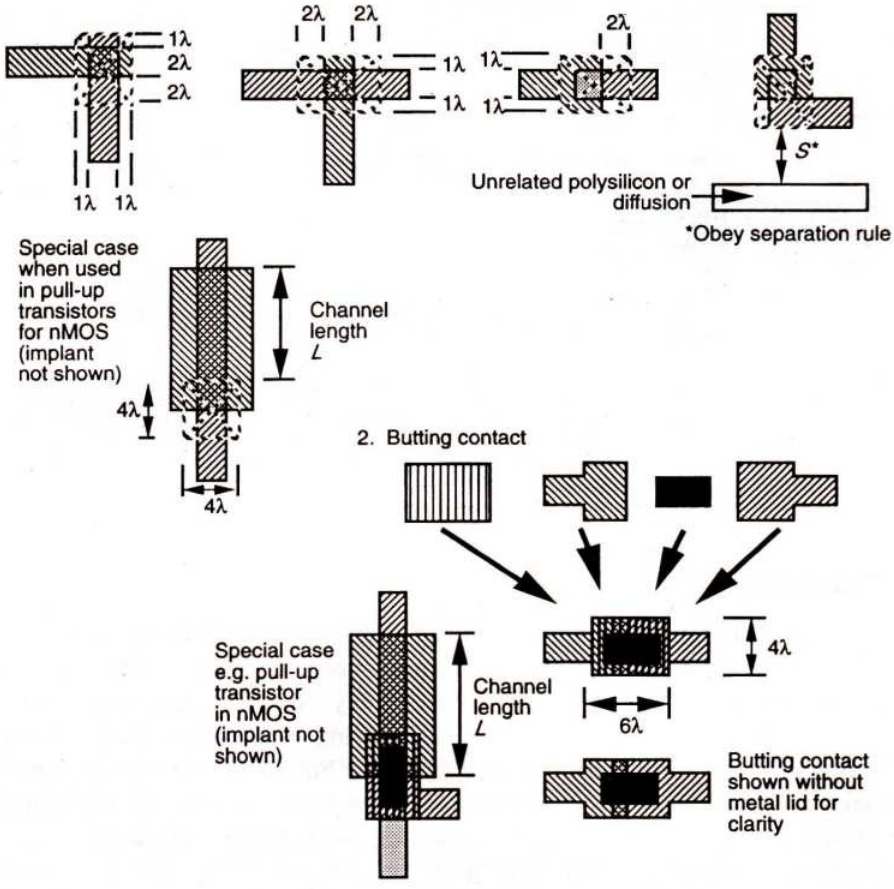


FIGURE 3.9 Contacts polysilicon to diffusion (nMOS only in the main text).

The 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

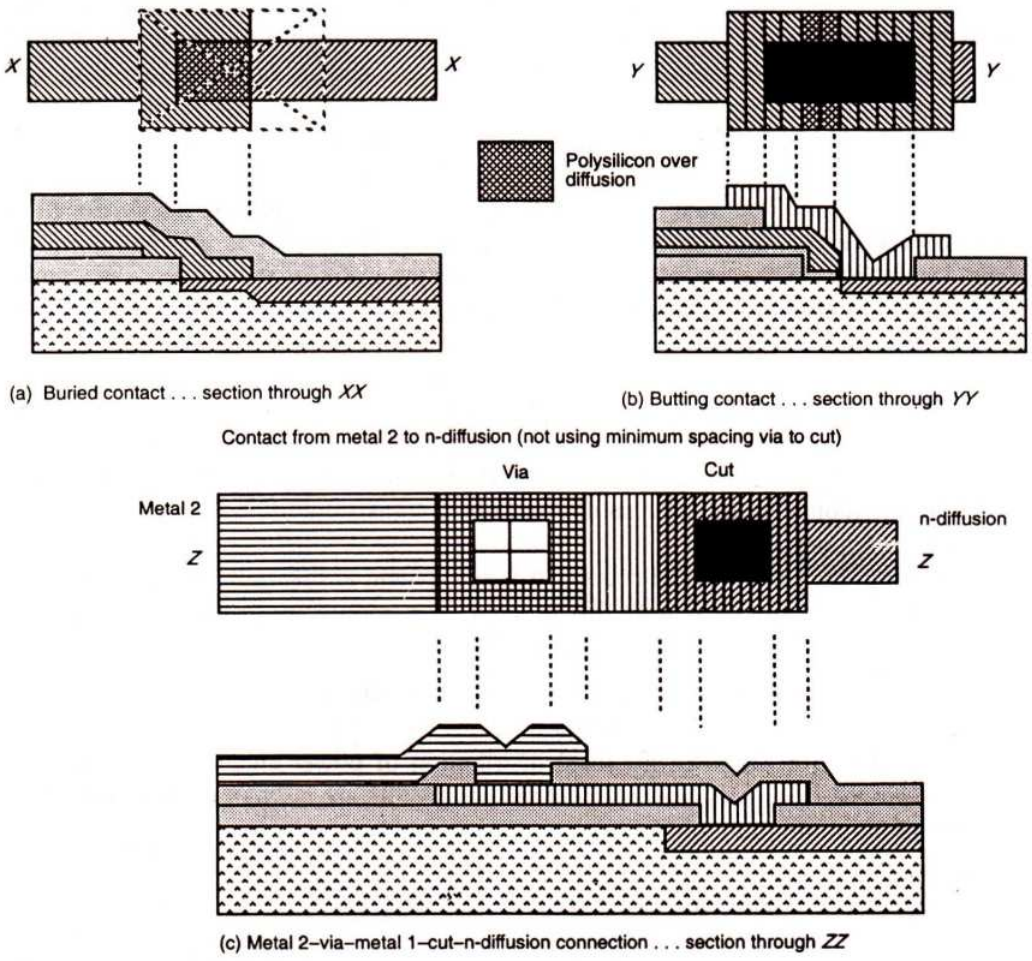


FIGURE 3.10 Cross-sections through some contact structures.

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 to be 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.
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 nMOS 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.

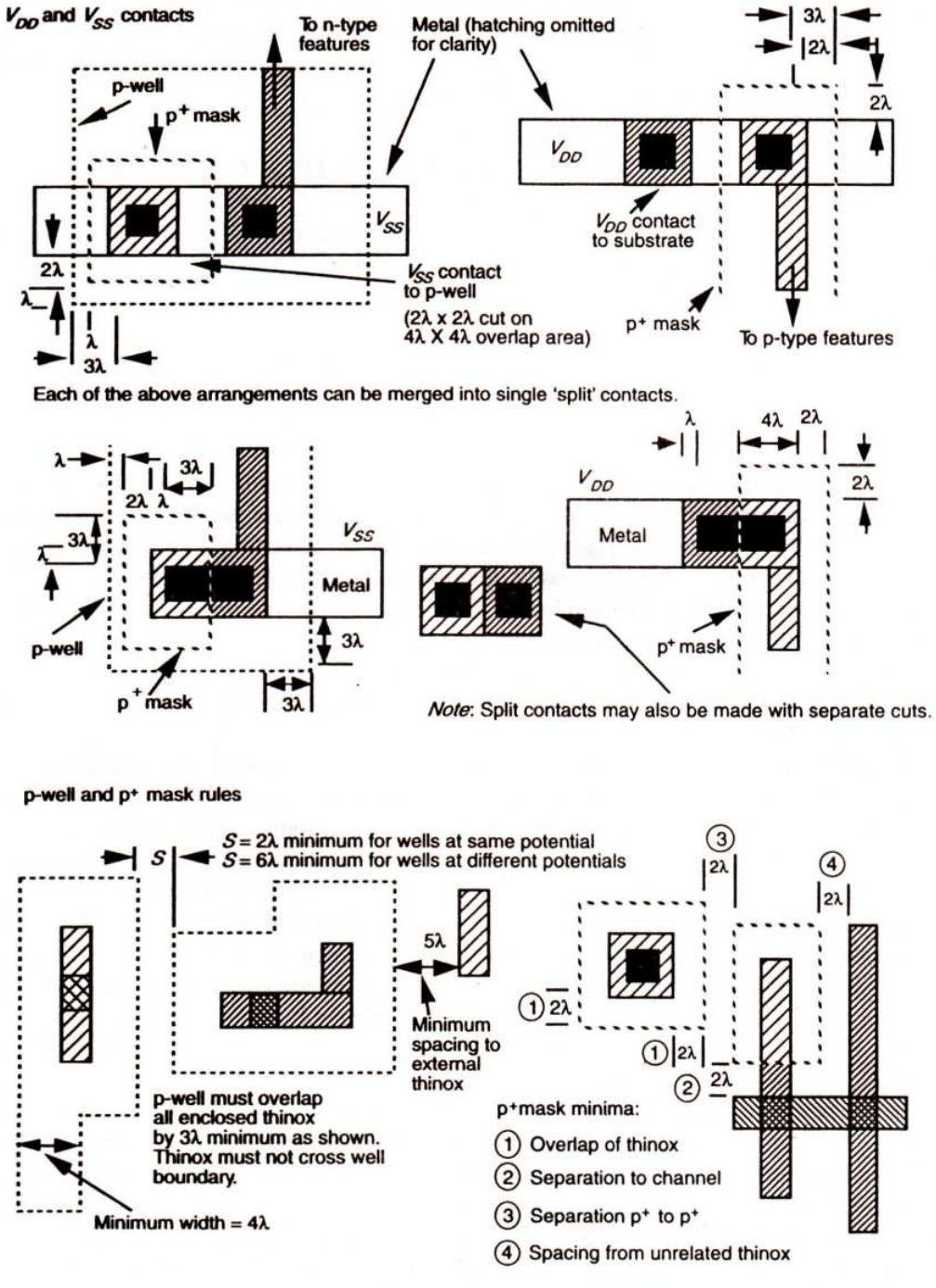


FIGURE 3.11 Particular rules for p-well CMOS process.

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 λ which is related to the resolution of the process. λ 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 2λ in their interrelationship. Inevitably, a consequence of using the lambda-based concept is that every dimension must be rounded up to whole λ 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.

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.

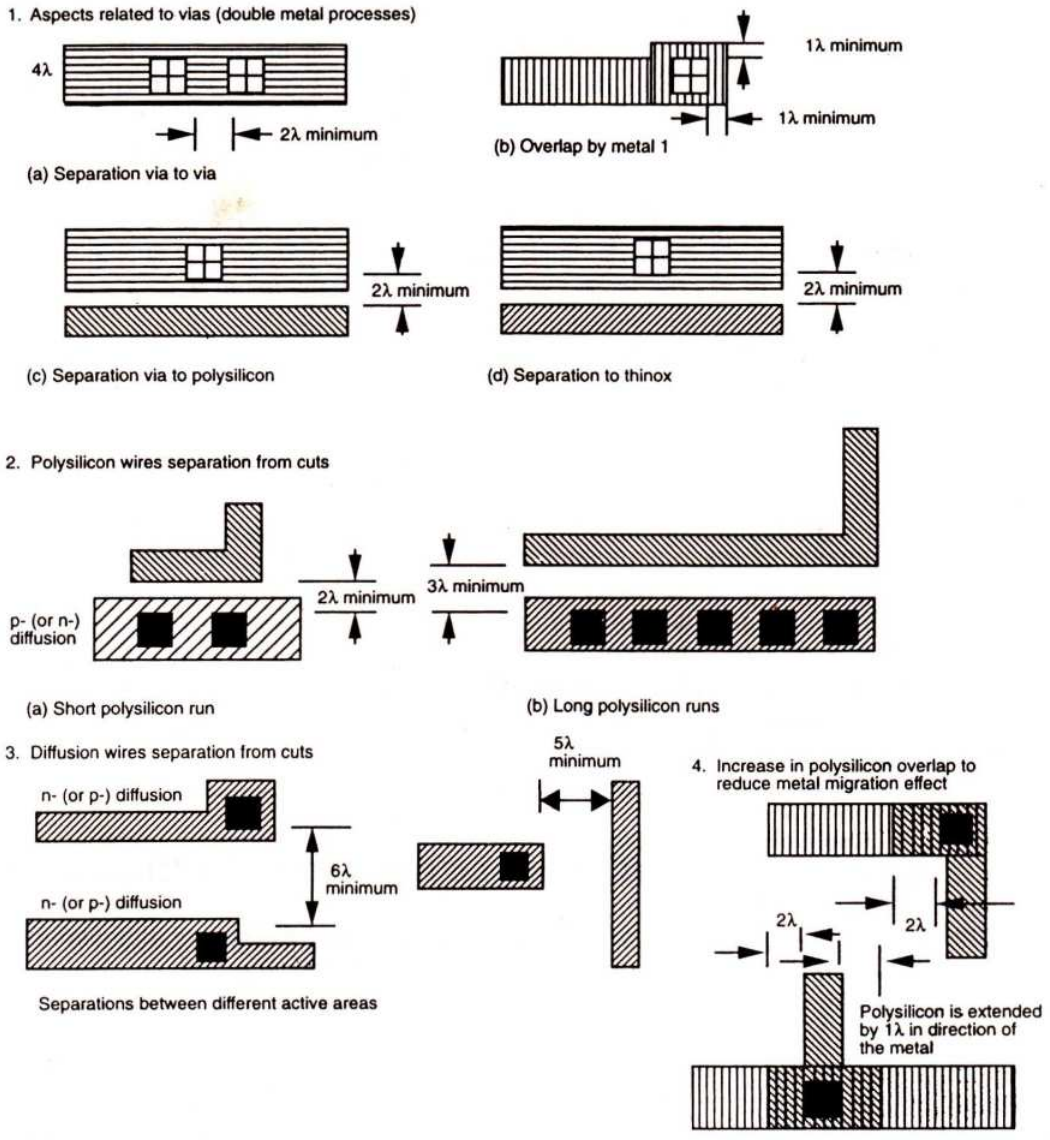


FIGURE 3.12 Further aspects of λ-based design rules for contacts, including some factors contributing to higher yield/reliability.

The 3λ 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 3λ. In effect, this implies that the minimum feature size for thick oxide is 3λ.

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.

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 μm feature size double metal, double poly., n-well CMOS rules which also offer a BiCMOS capability.

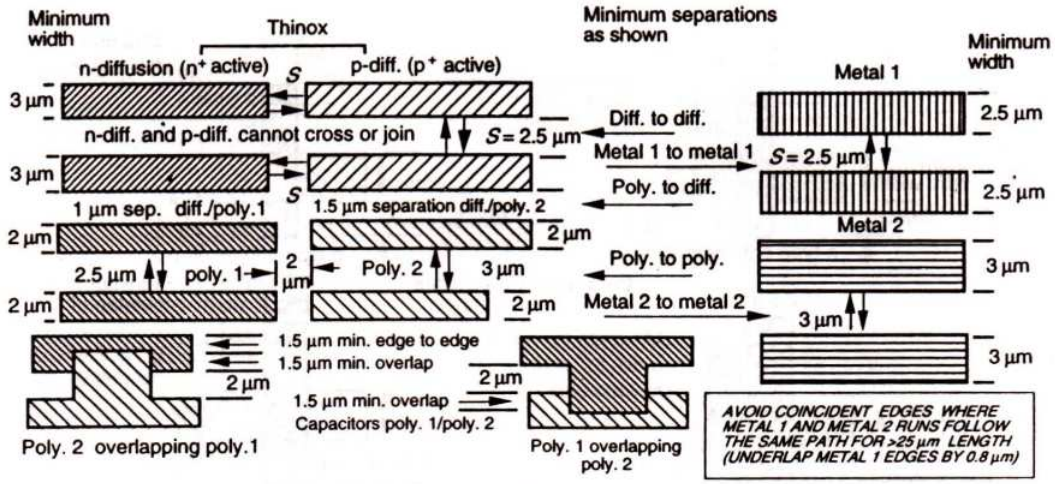
3.5 2 μM DOUBLE METAL, DOUBLE POLY. CMOS/BICMOS RULES*

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^+ subcollector—pale green; p-base—pink. These extra features are set out in Figure 3.1(c) and in Color plate l(c).

The use of color encoding is illustrated in the Color plates section of this book. The monochrome encoded rule set for the Orbit™ 2 μm 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

* 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 µm CMOS).

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 µm process are included in Appendix A.

3.6 1.2 µM 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 µm 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 λ and, for micron-based rules, it will be taken to represent the least common factor associated with the rules (for example, 0.25 µm per side for the 2 µm process and 0.2 µm per side for the 1.2 µm Orbit™ process layout). Most CAD VLSI tools also offer convenient facilities for mask level design.

* 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.

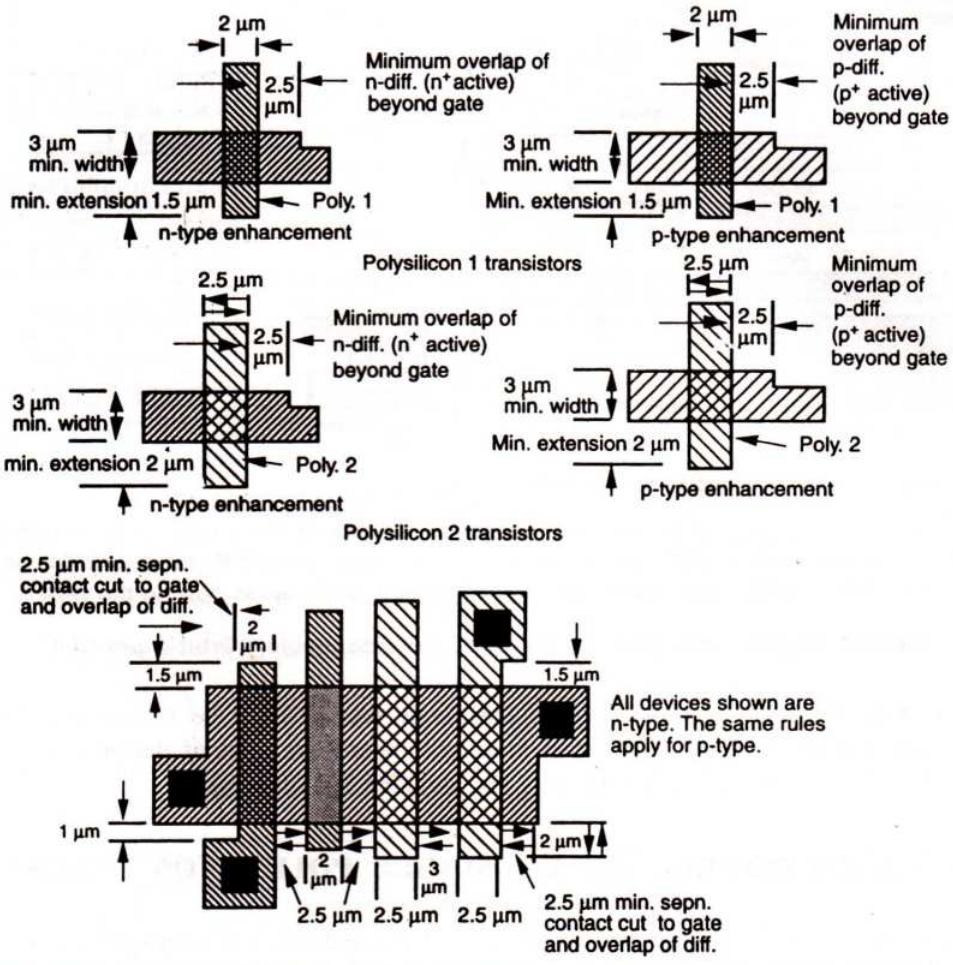


FIGURE 3.13(b) Transistor related design rules (Orbit 2 µm CMOS) minimum sizes and overlaps.

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.

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).

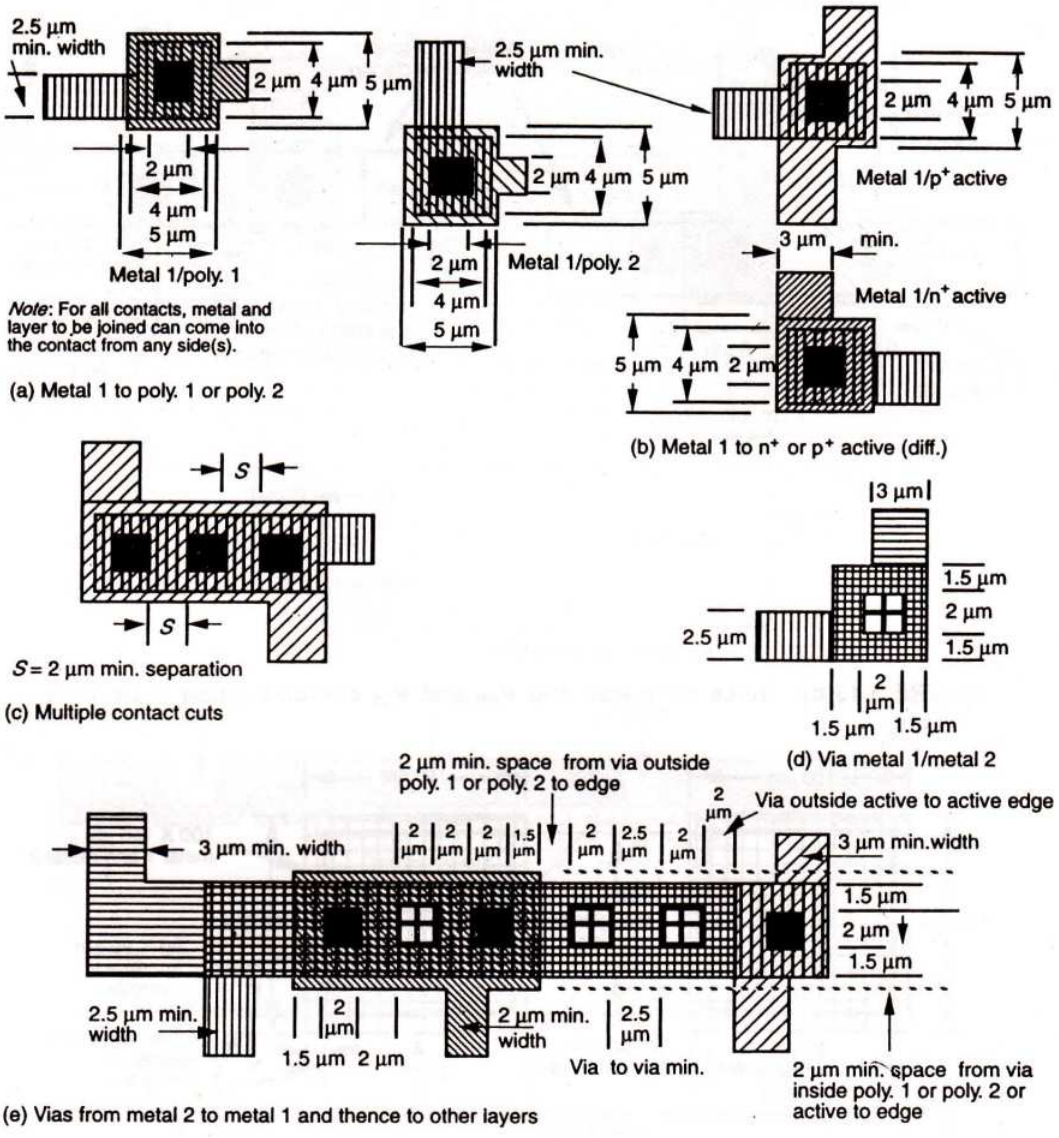


FIGURE 3.13(c) Rules for contacts and vias (Orbit 2 μm CMOS).

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.

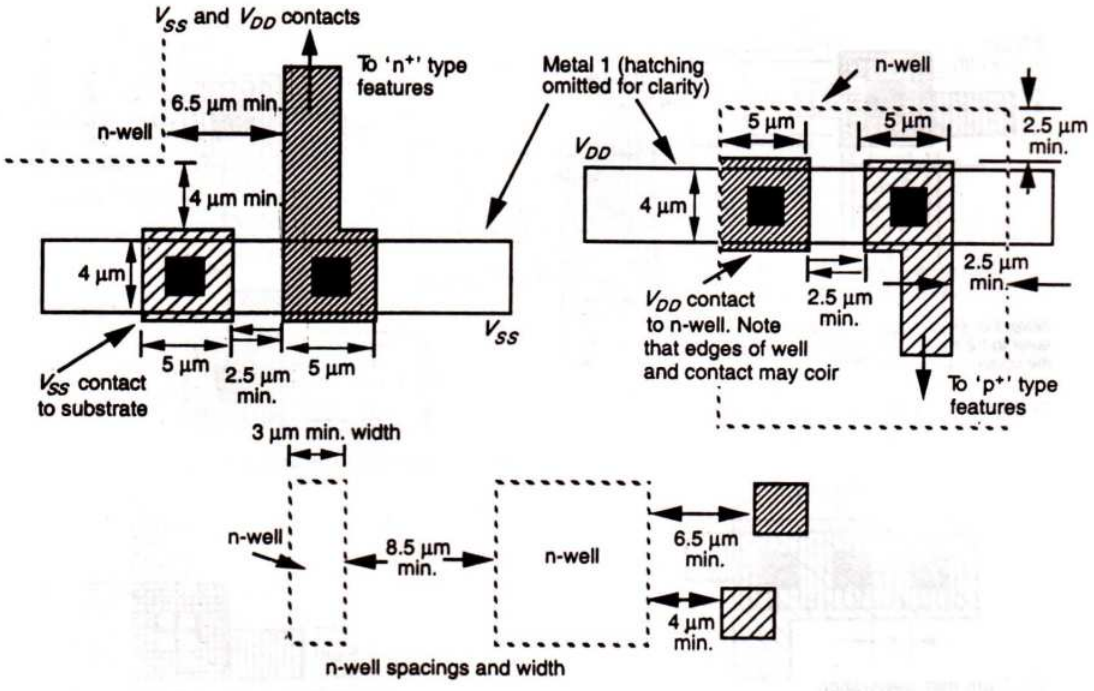
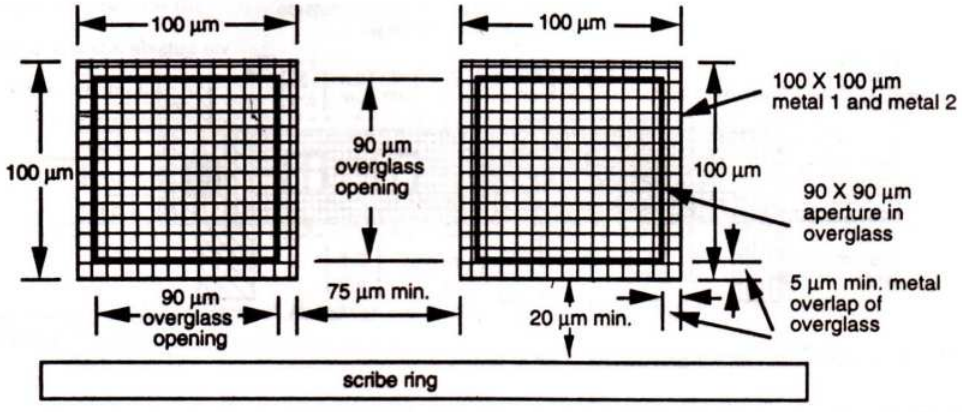
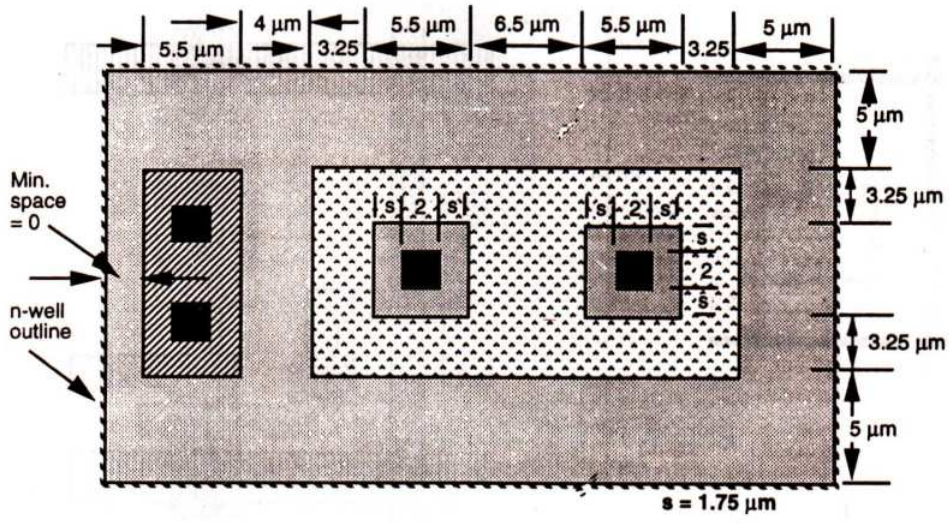


FIGURE 3.13(d) Rules for n-well and V_{DD} and V_{SS} contacts (Orbit 2 µm CMOS).

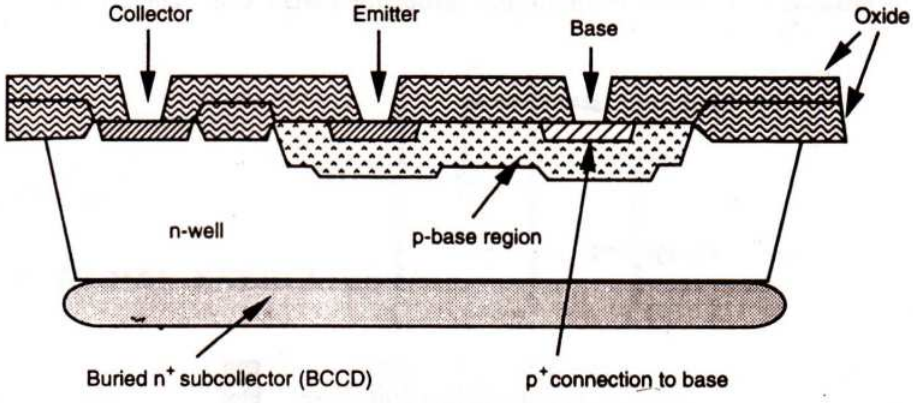


Other rules and encodings:
Via overlap of pad 2 µm.
Pad to active separation 20 µm min.
Color encoding for overglass mask . . . Gray

FIGURE 3.13(e) Rules for pad and overglass geometry (Orbit 2 µm 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 µm BiCMOS)

FIGURE 3.13(f) Special rules for BiCMOS transistors (Orbit 2 µm CMOS).

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 µm 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

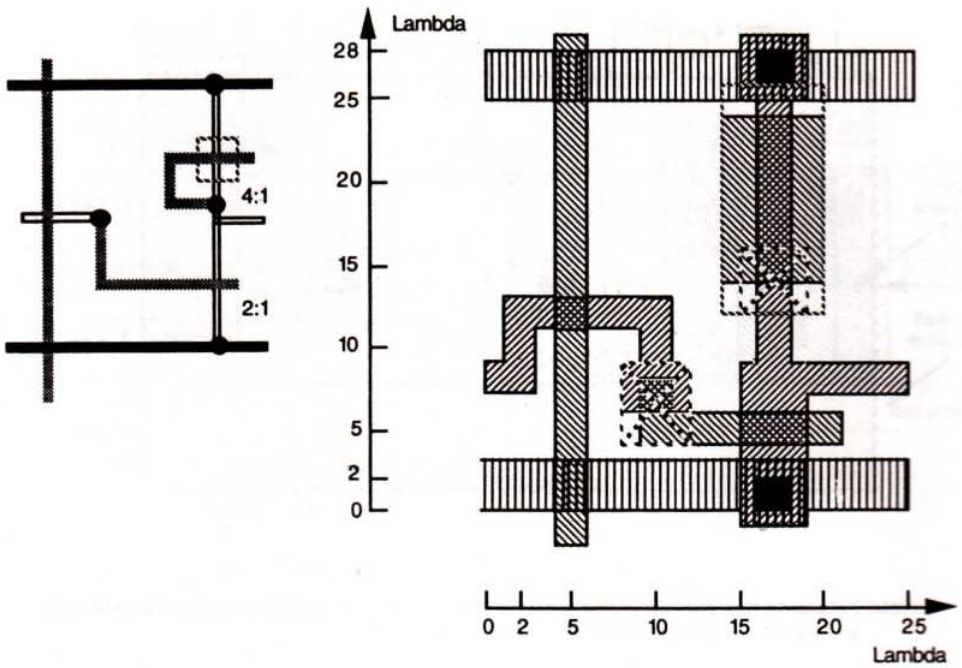


FIGURE 3.14 Stick diagram and layout for nMOS shift register cell.

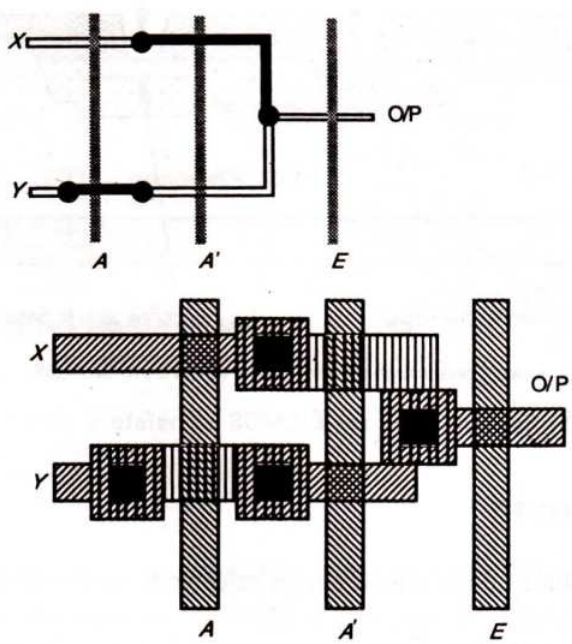


FIGURE 3.15 Two-way selector with enable.

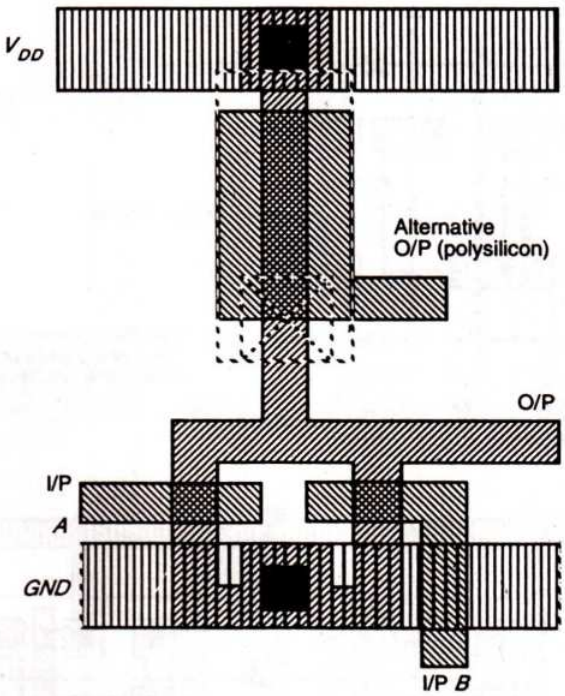


FIGURE 3.16 Two I/P nMOS *Nor* gate.

the budding designer becomes 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.

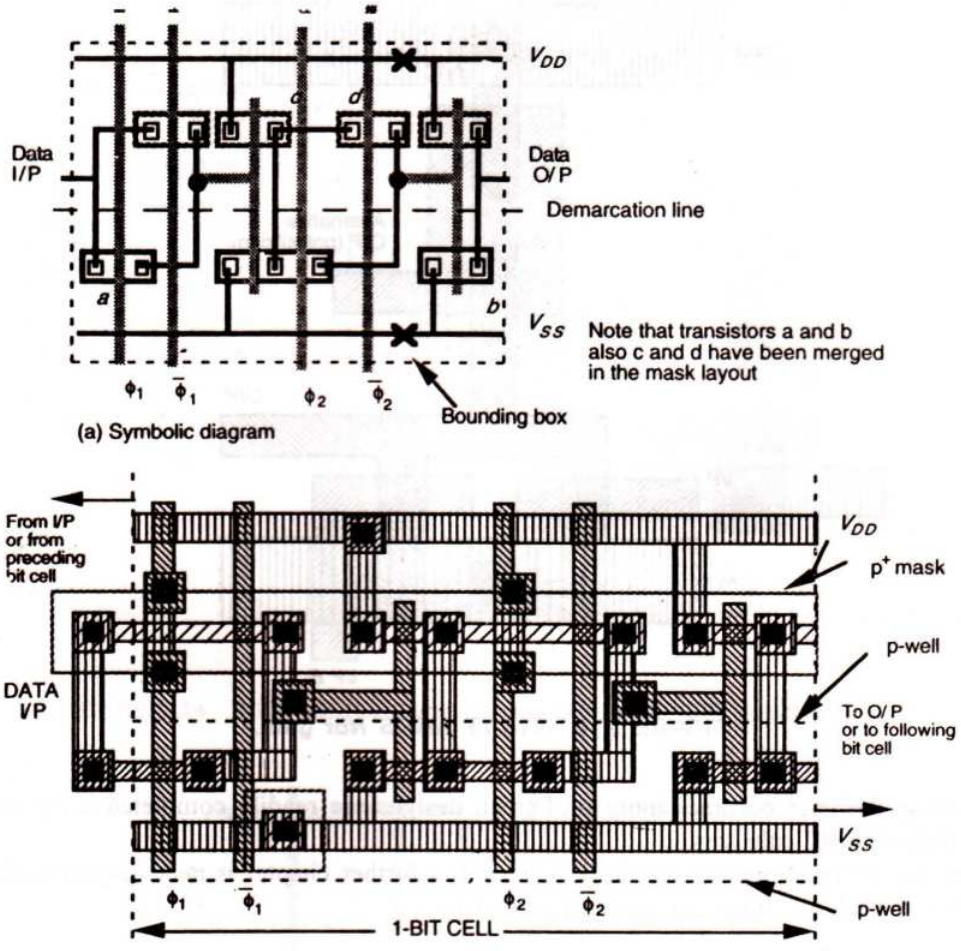


FIGURE 3.17 A 1-bit CMOS shift register cell.

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.

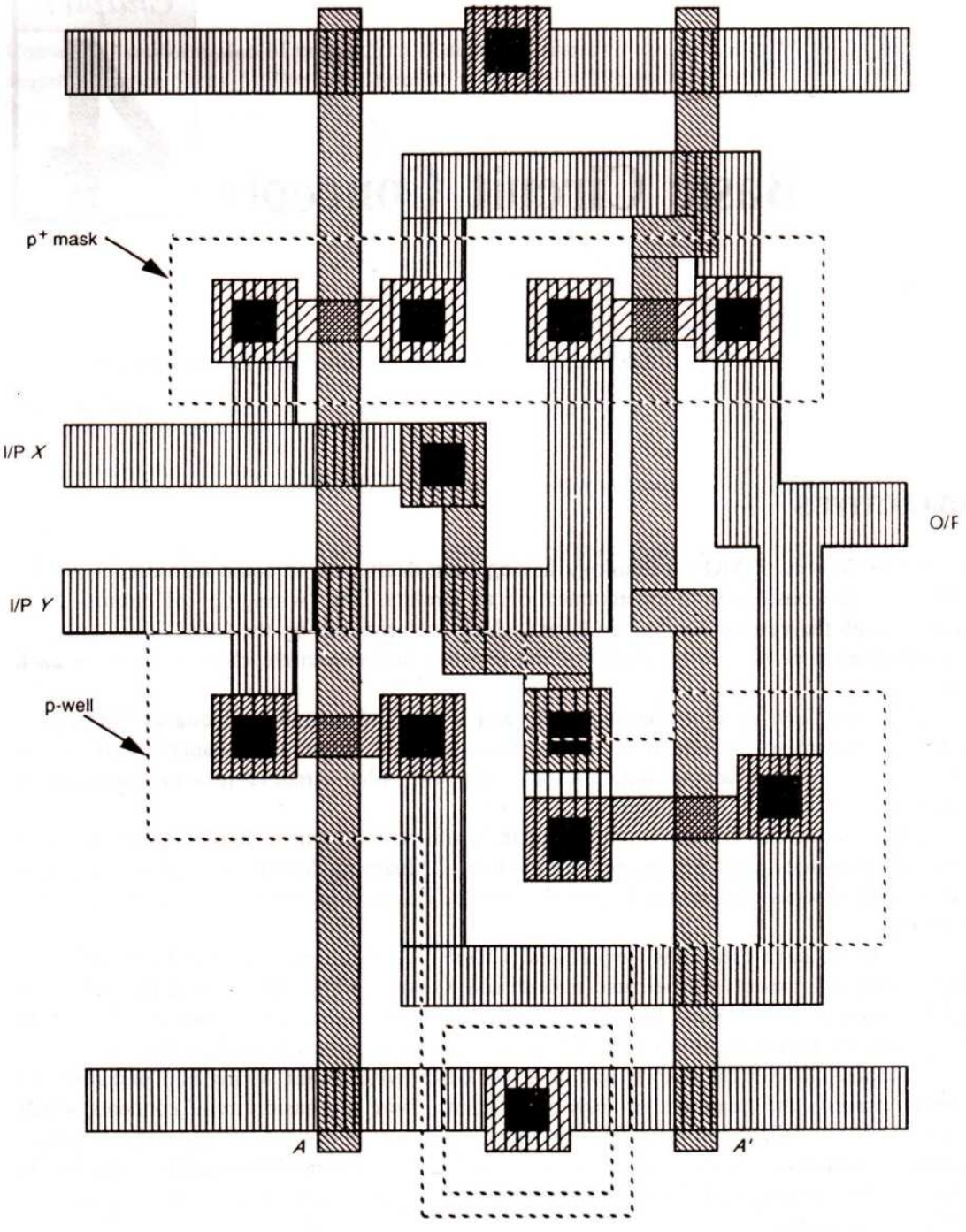


FIGURE 3.18 CMOS layout example.

Chapter 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 $\square Cg$, help greatly in evaluating the effects of wiring and input and output capacitances. Further, the delays associated with wiring, with inverters and with other circuitry may be conveniently evaluated in terms of a delay unit τ.

Parameter values for the layers in 5 μm, 2 μm and 1.2 μm 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 ρ, of width W, thickness t, and length between faces L. The arrangement is shown in Figure 4.1.

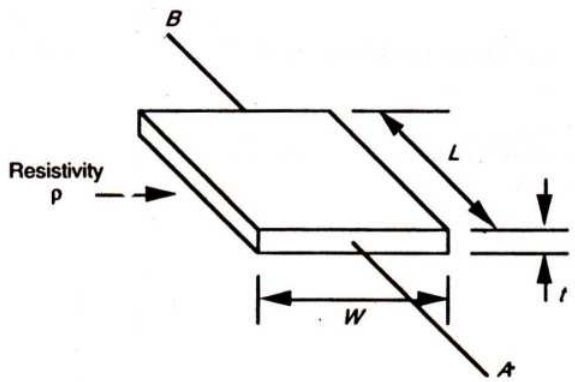


FIGURE 4.1 Sheet resistance model.

With reference to Figure 4.1, consider the resistance R_{AB} between two opposite faces.

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

where

A = cross-section area

Thus

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

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

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

where

R_s = ohm per square or sheet resistance

Thus

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

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

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

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

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

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

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

* 5 micron (μm) technology implies minimum line width (and feature size) of 5 μm and in consequence λ = 2.5 μm. Similarly, 2 μm and 1.2 μm technologies have feature sizes of 2 μm and 1.2 μm 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 I_{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.

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

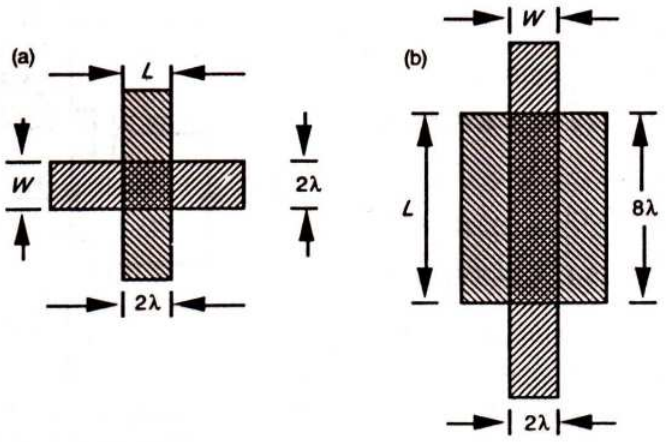


FIGURE 4.2 Resistance calculation for transistor channels.

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

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

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

Thus, channel resistance

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

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

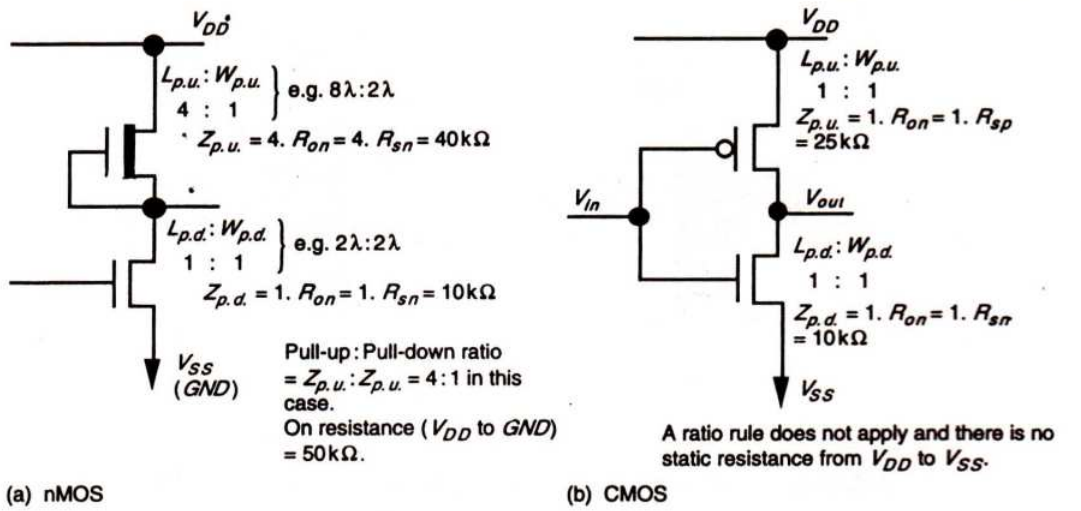
Figure 4.3 takes these considerations one step further and shows how the pull-up to pull-down ratio of an inverter is determined. In the nMOS case a simple 4:1 $Z_{p.u.}$:$Z_{p.d.}$ ratio obviously applies. Note, for example, that a 4:1 ratio would also be achieved if the upper channel (p.u.) length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

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

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

* From Table 4.1.



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

FIGURE 4.3 Inverter resistance calculation.

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:

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

where

D = thickness of silicon dioxide

A = area of plates

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

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

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

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

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

(D in cm, ε_0 in farads/cm)

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

TABLE 4.2 Typical area capacitance values for MOS circuits

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

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 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 μm MOS circuits:

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

or, for 2 μm MOS circuits (Orbit):

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

and, for 1.2 μm MOS circuits (Orbit):

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

4.5 SOME AREA CAPACITANCE CALCULATIONS

The approach will be demonstrated using λ-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 (2λ × 2λ for λ-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.

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

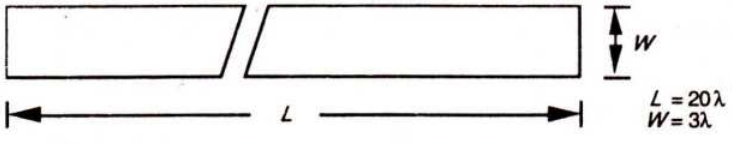


FIGURE 4.4 Simple area for capacitance calculation.

Now:

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

2. Consider the same area in polysilicon.

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

   $= 1.5 \square C_g$

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

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

   $= 3.75 \square C_g$*

Calculations of area capacitance values associated with structures occupying more than one layer, as in Figure 4.5, are equally straightforward.

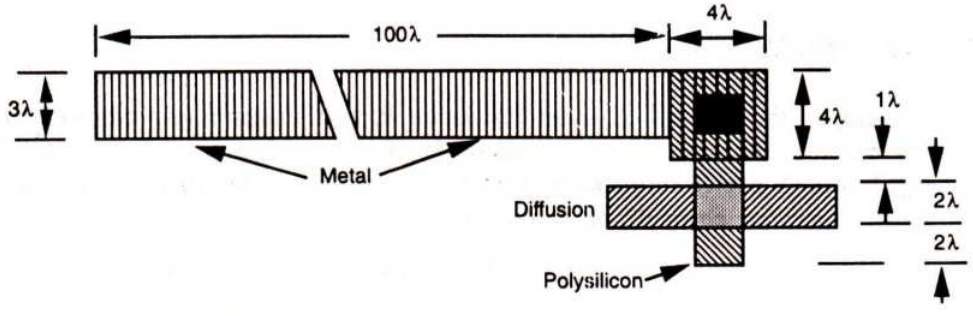


FIGURE 4.5 Capacitance calculation (multilayer).

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

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

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

Consider the polysilicon area (excluding the gate region)

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

Therefore

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

For the transistor,

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

* 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.

Therefore

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

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 4λ width but when taking this approach in design it must be borne in mind that, compared with 3λ width paths, the capacitance will be increased by one-third.

For example, if the metal width is increased to 4λ 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 τ

We have developed the concept of sheet resistance R_s and standard gate capacitance unit $\square C_g$. If we consider the case of one standard (feature size square) gate area capacitance being charged through one feature size square of n channel resistance (that is, through R_s for an nMOS pass transistor channel), as in Figure 4.6, we have:

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

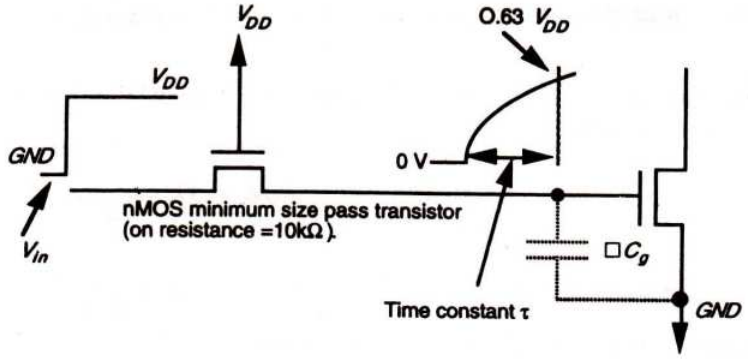


FIGURE 4.6 Model for derivation of τ.

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

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

and for 2 μm (Orbit) technology,

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

and for 1.2 μm (Orbit) technology,

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

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

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

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

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

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

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

$$= 0.13 \text{ nsec}$$

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

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

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

For 5 μm MOS technology τ = 0.3 nsec is a very safe figure to use; and, for 2 μm Orbit MOS technology, τ = 0.2 nsec is an equally safe figure to use; and, for 1.2 μm Orbit MOS technology, τ = 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 kΩ 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 τ = 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

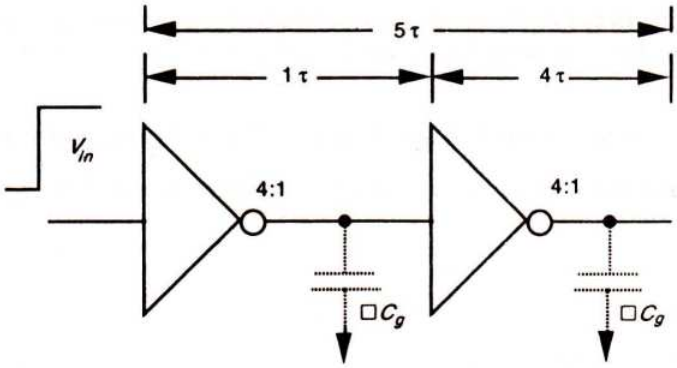


FIGURE 4.7 nMOS inverter pair delay.

through a pair of similar nMOS inverters is

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

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

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

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

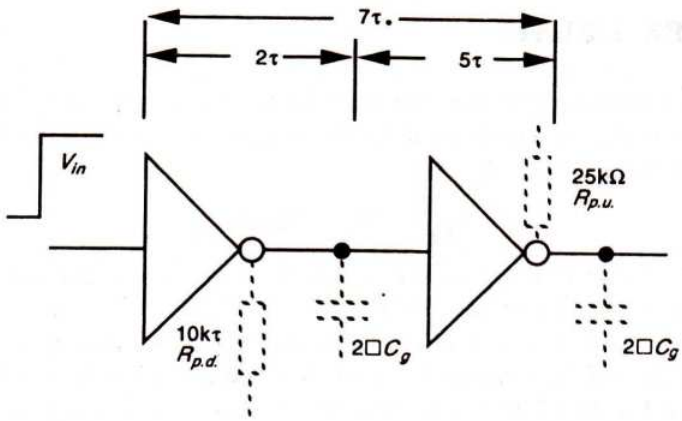


FIGURE 4.8 Minimum size CMOS inverter pair delay.

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

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 τ_r or fall-time τ_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

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

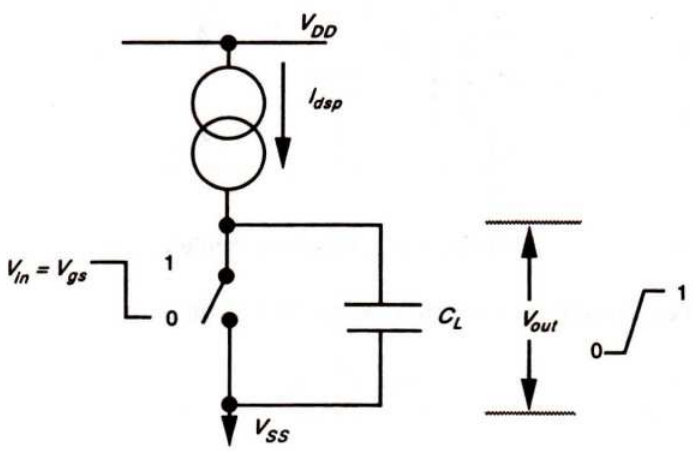


FIGURE 4.9 Rise-time model.

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

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

Substituting for I_{dsp} and rearranging we have

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

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

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

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

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

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.

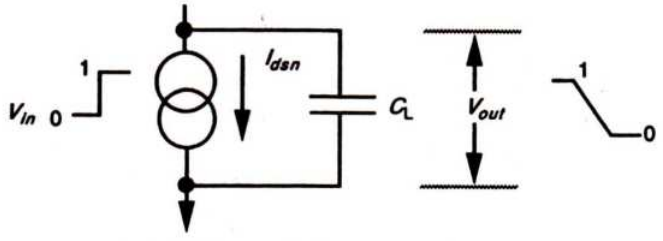


FIGURE 4.10 Fall-time model.

Making similar assumptions we may write for fall-time:

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

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

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

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

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

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).

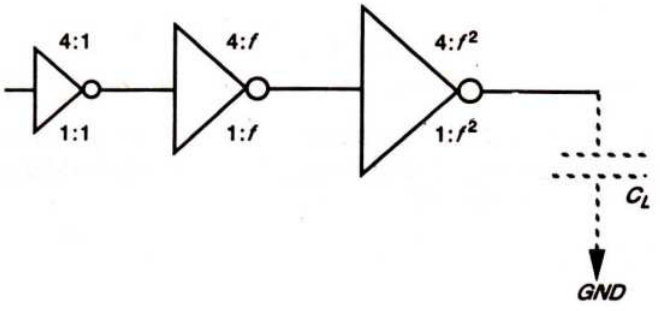


FIGURE 4.11 Driving large capacitive loads.

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

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

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

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

so that the choice of f and N are interdependent.

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

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

That is

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

Thus, for N even

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

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

Thus, in all cases

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

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

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

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

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

and overall delay t_d

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

or

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

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.

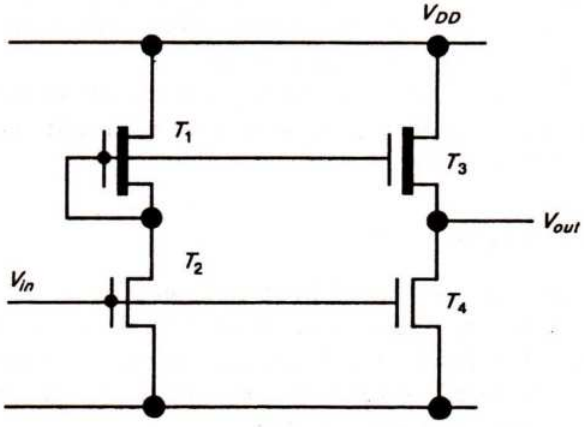


FIGURE 4.12 Inverting type nMOS super buffer.

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

Now consider the opposite transition: when V_{in} drops to 0 volt, then the gate of T_3 is allowed to rise quickly to V_{DD}. Thus, as T_4 is also turned off by V_{in}, T_3 is made to conduct

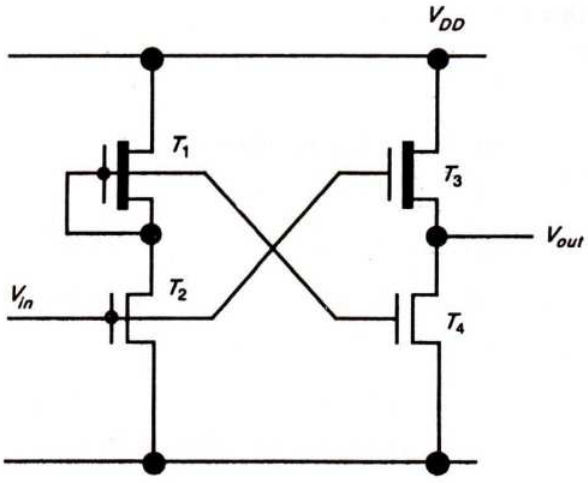


FIGURE 4.13 Non-inverting type nMOS super buffer.

with V_{DD} on its gate, that is, with twice the average voltage that would apply if the gate was tied to the source as in the conventional nMOS inverter. Now, since I_{ds} α V_{gs} then doubling the effective V_{gs} will increase the current and thus reduce the delay in charging any capacitance on the output, so that more symmetrical transitions are achieved.

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

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

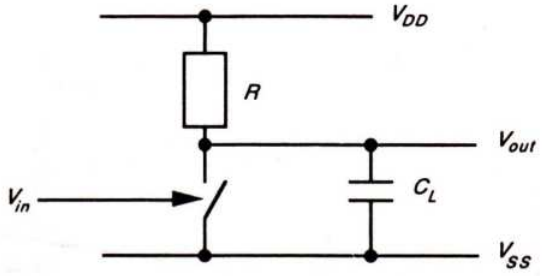
The availability of bipolar transistors in BiCMOS technology presents the possibility of using bipolar transistor drivers as the output stage of inverter and logic gate circuits. We have already seen (Chapter 2) that bipolar transistors have transconductance g_m and current/area I/A characteristics that 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 μ_n and collector current I_c it is only linearly dependent on T. This means that there is no difficulty in matching V_{be} values across a circuit, spread over an area on chip,

as the temperature differences across a chip will not be sufficient to cause more than a few millivolts of difference in V_{be} between any two bipolar transistors.

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



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.

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

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

where g_m is the transconductance of the bipolar transistor.

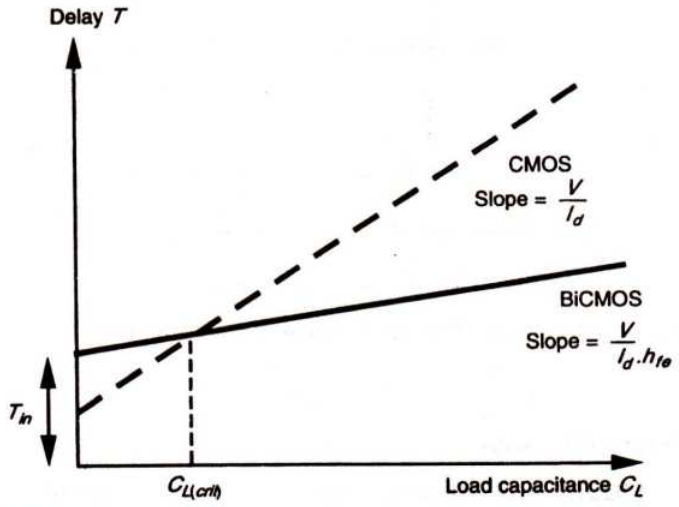
Clearly, since the bipolar transistor has a relatively high transconductance, the value of Δt is small.

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

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

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

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



- Delay of BiCMOS inverter can be described by

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

where

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

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

FIGURE 4.15 Delay estimation.

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

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

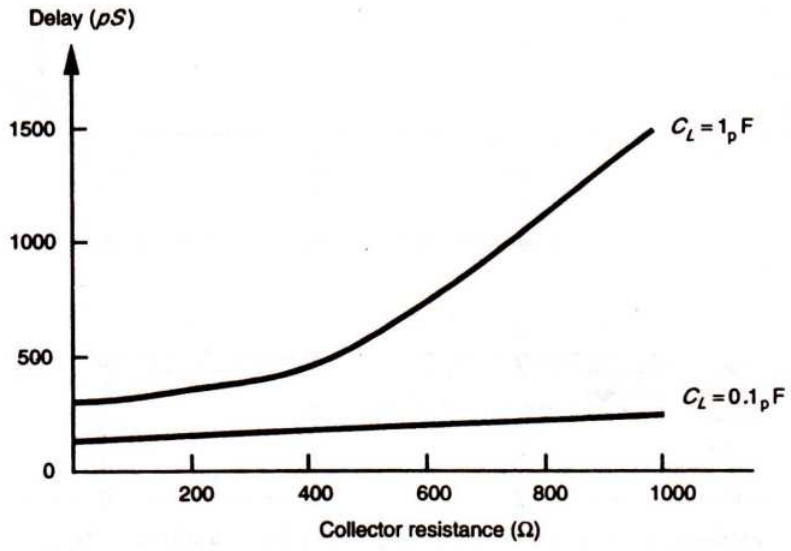


FIGURE 4.16 Gate delay as a function of collector resistance.

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

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

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

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

where

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

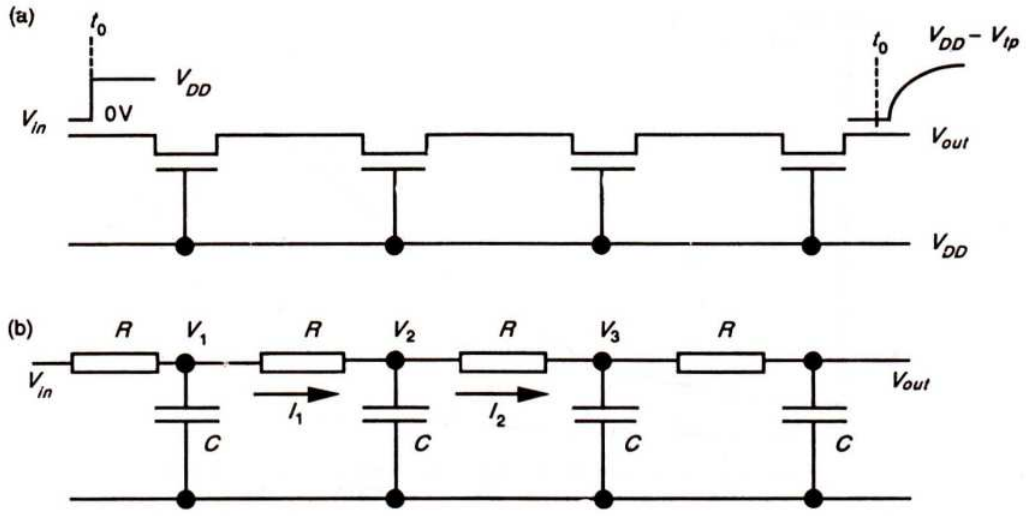


FIGURE 4.17 Propagation delays in pass transistor chain.

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

$$\tau_p \alpha x^2$$

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

$$R_{total} = nrR_s$$

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

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

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

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

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

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 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.

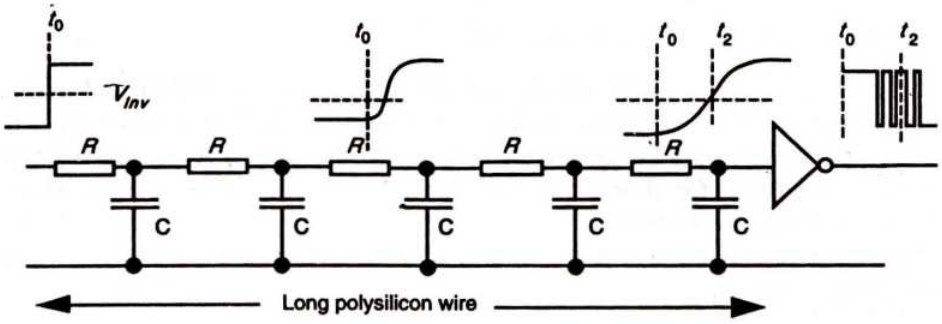


FIGURE 4.18 Possible effects of delays in long polysilicon wires.

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

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.

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

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 μm	*2 μm*	*1.2 μm*
Area C(C_{area}) (as in Table 4.2)	1.0×10^{-4} pF/μm^2	1.75×10^{-4} pF/μm^2	3.75×10^{-4} pF/μm^2
Periphery (C_{periph})	8.0×10^{-4} pF/μm	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 μm)	*μm-based (2 μm)*	*μm-based (1.2 μm)*
Metal	chip wide	chip wide	chip wide
Silicide	2,000λ	NA	NA
Polysilicon	200λ	400 μm	250 μm
Diffusion (active)	20λ*	100 μm	60 μm

* 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 $\rho = 1$ ohm cm. A section of this layer is 55 μm long and 5 μm wide and has a thickness of 1 μm. 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 3λ wide metal path which crosses a 2λ wide polysilicon path at right angles. Assuming that the layers

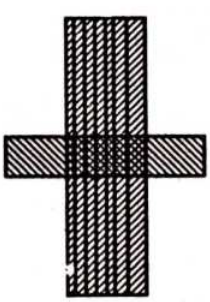


FIGURE 4.19 Layout detail for Question 2.

are separated by a 0.5 μm thick layer of silicon dioxide, find the capacitance between the two layers.

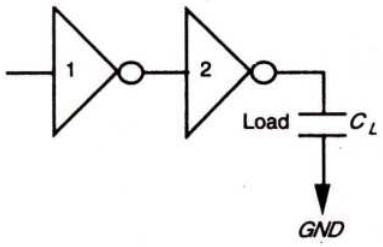
The polysilicon layer in turn crosses a 4λ wide diffusion region at right angles to form a transistor. Using the tables provided in the text, find the gate to channel capacitance. Compare it with the metal to polysilicon capacitance already calculated.

Assume λ = 2.5 μm 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 τ 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 τ and evaluate this pair delay.



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

FIGURE 4.20 Circuit for Question 3.

4. An off chip capacitance load of 5 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 pF load.
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.

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

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

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

$$= .05625\ pF$$

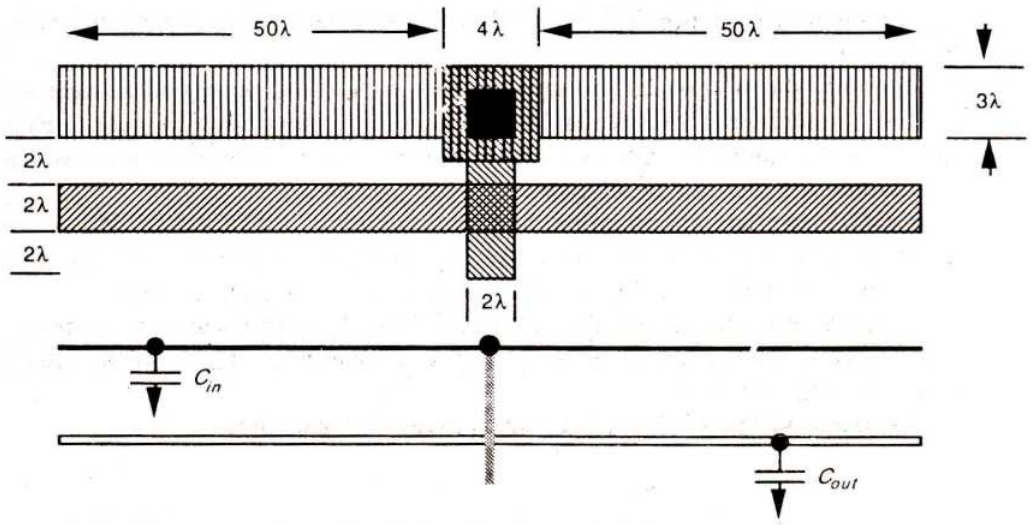


FIGURE 4.21 Structure for Question 5.

$$C_p = [(4 \times 4 + 2 \times 2 + 2 \times 1)\lambda^2 \times 6.25\ \mu m^2/\lambda^2]\ \{0.4 \times 10^{-4}\ pF/\mu m^2\}$$
$$= .0055\ pF$$
$$C_g = 1\square C_g = .01\ pF$$

Thus

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

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

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