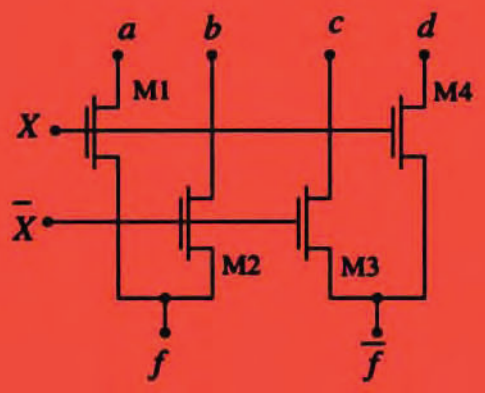
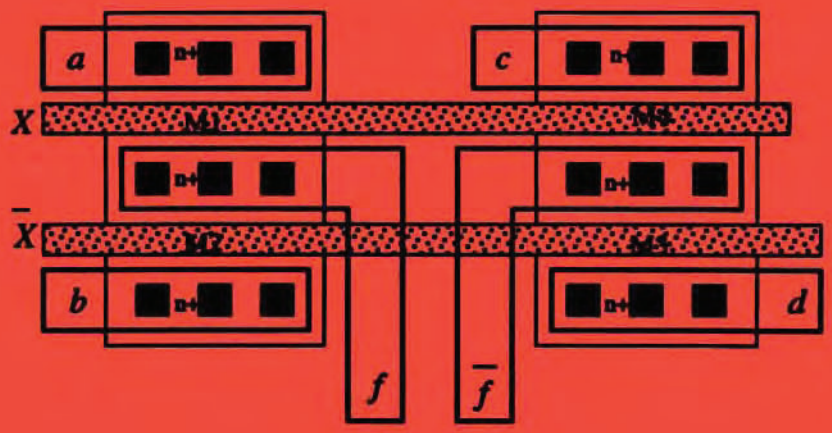
CMOS LOGIC CIRCUIT DESIGN





John P. Uyemura

CMOS LOGIC CIRCUIT DESIGN

John P. Uyemura
Georgia Institute of Technology

KLUWER ACADEMIC PUBLISHERS
NEW YORK, BOSTON, DORDRECHT, LONDON, MOSCOW

eBook ISBN: 0-306-47529-4
Print ISBN: 0-7923-8452-0

©2002 Kluwer Academic Publishers
New York, Boston, Dordrecht, London, Moscow

Print ©2001 Kluwer Academic Publishers
Dordrecht

All rights reserved

No part of this eBook may be reproduced or transmitted in any form or by any means, electronic, mechanical, recording, or otherwise, without written consent from the Publisher

Created in the United States of America

Visit Kluwer Online at: http://kluweronline.com
and Kluwer's eBookstore at: http://ebooks.kluweronline.com

which gives varying characteristics as the temperature is varied. Polysilicon is of particular importance in CMOS since MOSFET gates generally consist of a deposited poly layer with a refractory metal (such as W or Pt) either on top of it, or mixed in during the deposition process. This combination is called a **silicide.**

Polycrystal silicon gains it name because it consists of many small regions crystal, called crystallites, instead of having a single crystal structure throughout (such as in a silicon wafer). This state is achieved by depositing silicon over an amorphous material such as silicon dioxide. Silicon atoms attempt to form crystals, but do not have a well-defined base to grow on. This results in the formation of the local crystal regions called **crystallites.** Polysilicon is used because it provides excellent coverage, has good adhesion properties to the silicon dioxide surface, and can be subjected to high temperature processing steps. One of the drawbacks of the material is that even heavily doped poly has a substantial resistance to current flow. Adding a refractory metal overcomes this problem, and allows poly lines to be used as interconnect wiring.

2.1.3 Doping and Ion Implantation

High-density VLSI circuits use ion implantation to create **doped** n and p regions in the silicon substrate. A doped region is simply a section of the silicon into which impurity atoms have been purposely added to alter the electrical properties. Arsenic (As) and phosphorus (P) are used for n-type regions in which there is an enhanced free electron concentration. Boron (B) is used to create p-type regions where the equilibrium density of positively charged holes is greater than that of the electrons. In VLSI fabrication, doped regions are most often created using the technique of ion **implantation.**

In the ion implantation process, dopant ions are accelerated to high energies and literally smashed into the silicon wafer as shown in Figure 2.3. Collisions between the ions and the silicon atoms and electrons eventually bring the ions to rest within the wafer. This stopping mechanism creates a lot of damage to the crystal that must be repaired; in addition, the ions must find their way to normal silicon lattice sites in order to act like substitutional impurities. Both objectives are achieved by heating the wafer in an **annealing step** where thermal energy is used to induce diffusion of the atoms.

A simple analytic approximation for the ion implanted doping density is given by the Gaussian expression

$$N_{ion}(x) = N_p \exp\left[-\left(\frac{x - R_p}{2(\Delta R_p)}\right)^2\right] \tag{2.7}$$

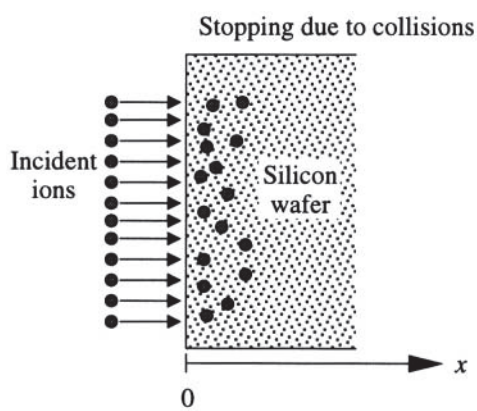


Figure 2.3 The ion stopping process

been doped to a given polarity (p-type or n-type) during the ingot growth. At the device level, nFETs require a p-type background, while pFETs are built in an n-type background. To create a complementary circuit that uses both types of transistors, we must provide an opposite polarity **well** in the process. This means that if a p-type substrate is used as the starting point, then nFETs can be fabricated directly in the substrate, but pFETs must reside in n-well regions that are added in a separate masking step.

The second important point that we need to consider is the fact that VLSI is based on the ability to achieve large transistor packing densities. For example, current commercial designs typically employ between 5 to 10 million MOSFETs or more on a single die. When the transistors are fabricated on the substrate, they must be electrically isolated from each other unless the circuit requires a connection. Isolation techniques are concerned with achieving this goal using a reasonably simple process that does not waste too much surface area.

As an example of the problem, let us examine the 3-transistor layout shown in Figure 2.12. The main idea is that we want to insure that there are no unwanted conduction paths between any two FETs. In particular, FET 1 and FET 2 are assumed to be isolated from each other, as are FET 2 and FET 3, using regions that are labelled "Isolation Oxide" which may be viewed as layers of glass that separate the individual transistors. If we want to establish an electrical connection between two devices, it must be done by means of a conducting layer. In the drawing, we have used a Metal1 layer to connect FET 1 to FET 3. Isolation is critical to the layout designer. It allows one to assume that adjacent transistors do not "talk" to each other unless an electrical connection is purposely added.

2.4.1 LOCOS

The most commonly used isolation technique is based on the **local oxidation of silicon (LOCOS).** This creates (relatively) thick silicon dioxide (quartz glass) insulating regions that surround every MOSFET. In standard terminology, we defined the substrate-level surface sections of the die as being either **Active** or **Field** such that

$$\text{Die} = \text{Active} + \text{Field}.$$

Active areas are flat regions on the silicon where MOSFETs reside; any region that is not Active is defined as Field.

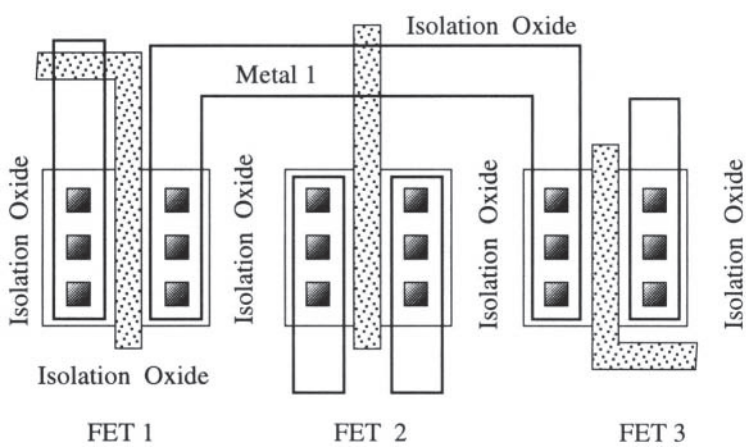


Figure 2.12 Use of isolation oxides to electrically separate FETs

Recall that the two reactions used to grow a thermal oxide layer on the silicon substrate are

$$Si + O_2 \rightarrow SiO_2$$

$$Si + 2H_2O \rightarrow SiO_2 + 2H_2 .$$

Both processes use silicon atoms from the substrate to create the SiO_2 layer with

$$x_{Si} \approx 0.46 x_{ox}$$

as derived earlier. LOCOS uses this fact to grow a recessed isolation oxide only in local field regions. LOCOS uses silicon nitride (Si_3N_4) to inhibit the growth of thermal oxides in Active areas of the die. The process is shown in Figure 2.13. The drawing in Figure 2.13(a) defines the starting point we have pre-determined the location of two adjacent nFETs. The drawing in (b) shows wafer after the initial first steps are completed. The substrate has been covered with a CVD silicon nitride layer that has been deposited on top of a thin thermal oxide layer. The underlying SiO_2 layer is called a **stress-relief oxide,** and is used to absorb the mechanical stress between the surfaces of the silicon wafer and the silicon nitride. If the stress-relief oxide is omitted, then the nitride[4] layer has a tendency to crack. An active device region is defined as a surface area that remains flat; Figure

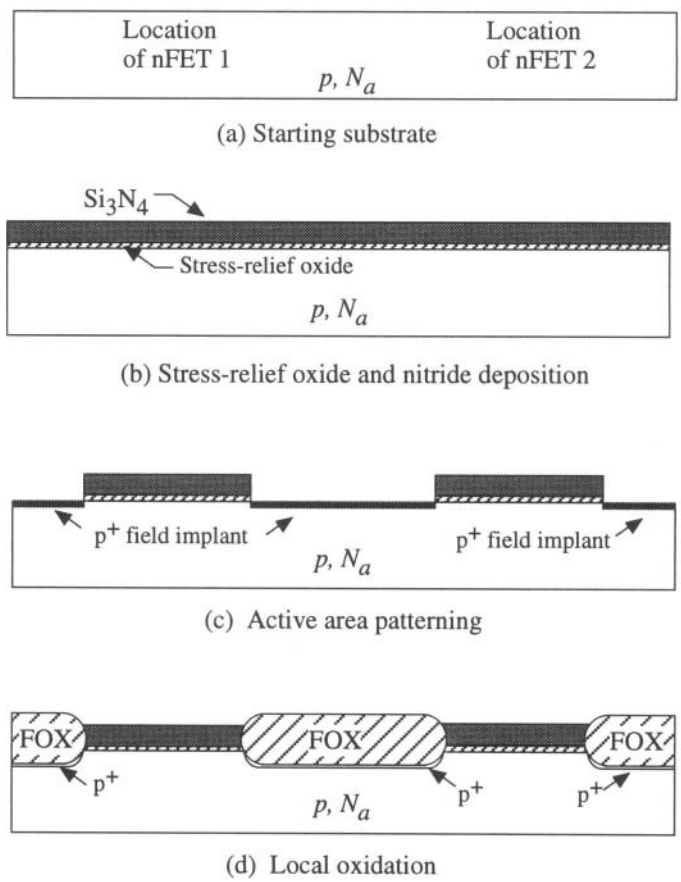


Figure 2.13 *Steps in the LOCOS isolation process*

[4] We use the term "nitride" to mean silicon nitride where there is no possibility of confusion.

2.13(c) shows the wafer after the nitride/oxide layer has been patterned such that the nitride remains in FET regions. Once the nitride has been patterned, boron ions are implanted into the exposed regions of the silicon wafer; this is called a **field implant** and is discussed in more detail below. The next step is the actual growth of the isolating **field oxide** (FOX). The surface of the wafer is subjected to an oxygen rich gas flow, producing thermal oxide in regions where the silicon is exposed, but protecting Active areas covered by the nitride. Due to the recession of the silicon surface, adjacent active areas are separated from each other by a recessed glass insulating region as shown in Figure 2.13(d). An alternate name for the field oxide is the **recessed oxide** (ROX) since it is indeed below the original surface of the silicon. The recession of the glass layer provides the desired electrical isolation between adjacent devices since it acts like a "wall of insulating glass" between them.

Growing the field oxide gives rise to the formation of an interface between the thick field oxide and the stress-relief oxide that is called the **bird's beak** region because of its shape. This is shown in Figure 2.14. Bird's beaking occurs because oxygen diffuses underneath the edges of the nitride, allowing oxide growth there. In creating this transition region, the edges of the nitride are lifted, reducing the size of the flat Active area. This phenomena is called **encroachment** (of the Active area) and reduces the integration density.

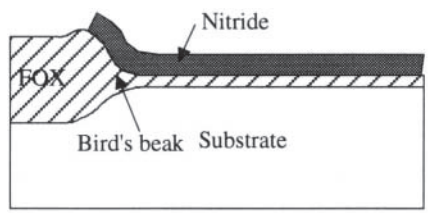


Figure 2.14 Bird's beak at nitride edge

Now let us examine the transistor isolation in more detail. Two adjacent FETs are portrayed in Figure 2.15. Since the FOX is recessed into the wafer surface, the glass acts to block direct electrical conduction between the two transistors. The drawing also shows the use of a metal interconnect line that is routed over the central field region. This creates a parasitic MOS capacitor structure that is characterized by the field oxide capacitance per unit area

$$C_{FOX} = \frac{\varepsilon_{ox}}{X_{FOX}} \tag{2. 27}$$

which is much smaller than C_{ox}. The field threshold voltage V_{TF} will be larger than the FET thresh-

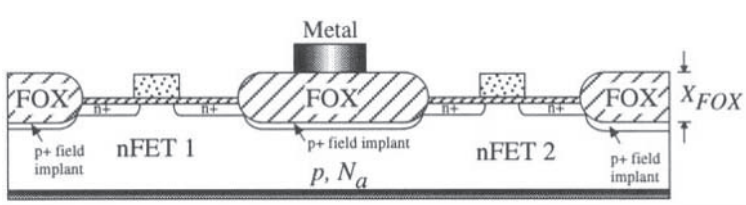


Figure 2.15 Side-view of adjacent nFETs isolated by a field oxide

old voltage V_{T0n}, but it is important to insure that the voltage on the interconnect is always less than V_{TF}. If the voltage exceeds this value, an inversion layer will form under the field oxide, and the isolation scheme fails. The field implant is used to adjust the field threshold voltage V_{TF} to a value that is much larger than normal operating voltages. It is also used for special input circuits that protect the internal circuitry from electrostatic discharge (ESD) problems; this is discussed in more detail in Chapter 10.

2.4.2 Improved LOCOS Process

Another problem with the simple process described above is that there is a difference in the height of Active and Field regions. This can be overcome by using an additional etching step as summarized in Figure 2.16. As seen in Figure 2.16(b), field regions of the silicon wafer are etched according to the patterning of the nitride layer. The depth is chosen to give a smooth surface after the field oxide is grown (or deposited, depending upon the process flow) as in (c). The final result helps maintain a flat surface as additional layers are added. CMP (chemical-mechanical polishing) techniques are very useful in this type of technology. Planarization has become increasingly important as the number of interconnect layers has increased.

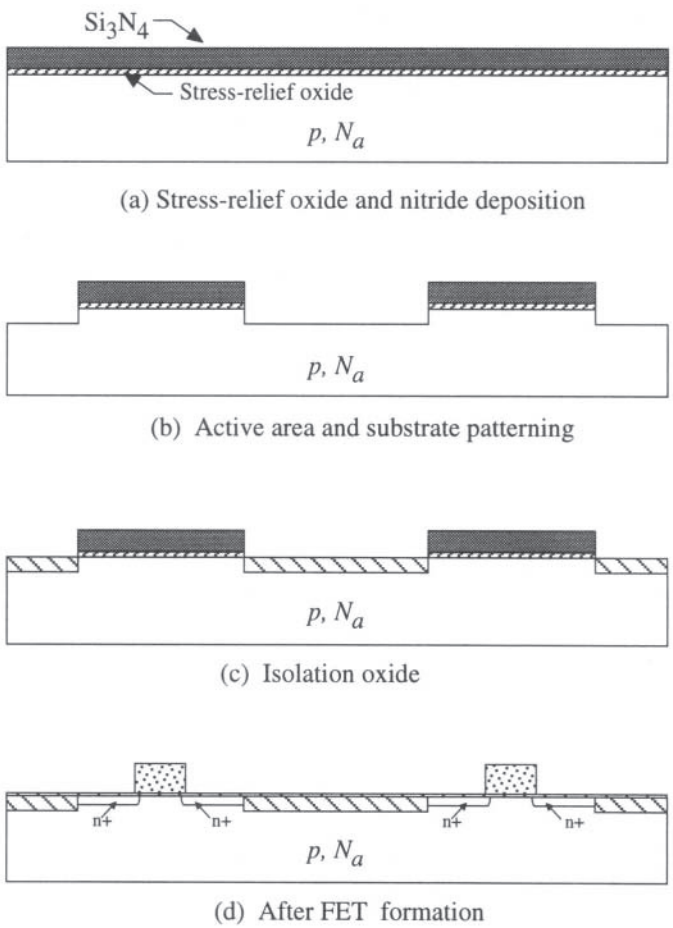


Figure 2.16 Planarized isolation technique

2.4.3 Trench Isolation

An alternate to LOCOS uses oxide-coated trenches that have been etched into the silicon wafer to isolate devices. Although the technology needed to perform trench isolation is more difficult than that used for LOCOS, the trenches require less surface area and thus allow for a higher integration density. In addition, some dynamic random-access memory (DRAM) cells use the trenches to create charge storage capacitors.

Conceptually, trench isolation is very straightforward to study. Reactive ion etching (RIE) is used to define vertical-walled trenches in the silicon substrate as shown in Figure 2.17(a). Oxygen is passed over the wafer, growing an oxide layer on the walls and bottom. Polysilicon is then used to fill-up the trenches, and a final oxide growth and addition of the FETs gives the structure shown in Figure 2.17(d). The isolation action is obvious: lateral current flows (i.e., parallel to the surface) are blocked by the insulating trenches. Overall, this yields a more planar surface that makes it easier to add above-wafer interconnect layers.

2.5 The CMOS Process Flow

Let us now follow the basic sequence that is needed to fabricate nFETs and pFETs in a p-type wafer. This is called an n-well process, since an n-region must be introduced to accommodate the pFETs. The cross-sectional views of the wafer during the steps described in the next few para-

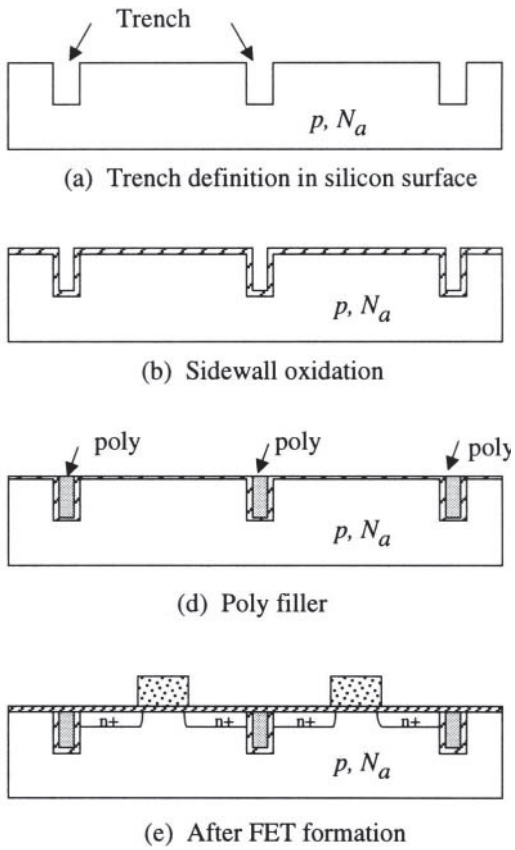


Figure 2.17 Basic steps in trench isolation

graphs are shown in the drawings of Figures 2.18 to 2.22. It is important to note that the drawings are not to scale since the thicknesses have been exaggerated for clarity.

The starting point in our example process is a heavily doped p^+ wafer that we will generally call the substrate. A thin p^- silicon epitaxial layer is grown on top of the wafer to provide a well-defined background for the transistors; this results in the cross-sectional view with the general structure depicted in Figure 2.18(a). Typically, the boron (acceptor) doping of the epitaxial layer is less than about N_a=10^{15} cm^3, and the epi layer is only a few microns thick. In the remaining views, only the epitaxial layer will be shown; the wafer itself will be omitted in an attempt to preserve at least some of the scaling. Since nFETs have a p-type bulk, they can be created in the epitaxial layer. A pFET, on the other hand, requires an n-type bulk, so that we provide an n-well in the p-epitaxial layer for these transistors. The n-well is created using a deep ion implant that is diffused deeper into the substrate, resulting in the cross-section shown in Figure 2.18(b). Once this has been accomplished, the threshold voltage ion implant adjustments must be performed for both nFETs and pFETs. These are shown in Figures 2.18(c) and (d). Creating a positive nFET threshold voltage V_{Tn} requires a boron (p-type) with a dose D_I determined by the oxide thickness, doping levels, and other physical parameters. The working value of the pFET threshold voltage $V_{Tp} < 0$ is also established by a ion implantation step; a donor implant makes the value more negative while an acceptor implant makes the value less negative. The next step is the creation of the dielectric isolation regions. This is summarized by the steps shown in Figure 2.18(e) and (f) for a LOCOS process, and results in the characteristics discussed in the previous section.

The next group of processing steps are used to form the transistors themselves. Access to the bare silicon surface [Figure 2.19(a)] is achieved by stripping the nitride and stress-relief oxide layers. This allows the careful growth of the gate oxide layer in which x_{ox} is established as in Figure 2.19(b). The gate oxide establishes the value of the oxide capacitance per unit area

$$C_{ox} = \frac{\varepsilon_{ox}}{x_{ox}} \tag{2. 28}$$

and is considered one of the most critical steps in the CMOS process flow. The next step is to deposit the gate polysilicon layer, which is then patterned by lithography according to the location of the transistor gates; this results in the structure shown in Figure 2.19(c).

The transistors themselves are formed by ion implants using the self-aligned scheme. pFETs are created using a p-type boron implant in which nFET locations are blocked with photoresist; the resulting cross-section is portrayed in Figure 2.19(d). Similarly, nFETs require an n-type implant for drain and source regions. To accomplish this, pFET locations are blocked with resist while the n-implant is performed, leaving the structure shown in Figure 2.19(e). At this point, both FET polarities are established.

Figure 2.20 illustrates the "above-wafer" steps in the processing sequence. In (a), the surface has been coated with a CVD oxide that acts to electrically insulate the device from overlaying conductors. In advanced processes, this is followed by steps that planarize the surface for the next material layer. Contact cuts are etched into the oxide where needed, and then filled with a metal "plug" such as tungsten [Figure 2.20(b)] . The first layer of metal interconnect, denoted generically as "Metal1" is deposited and patterned as implied by the view in Figure 2.20(c).This is repeated for each subsequent interconnect level. Figure 2.21 illustrates the cross-sectional view after the deposition of the second metal (denoted as metal2 in the drawing); in this case, the contacts are called "vias."

State-of-the-art CMOS processing has become quite complex, with many interconnect layers used to ease the layout and signal distribution problems. Figure 2.22 shows the layers in a hypothetical 4-metal process.The numerical values indicate thicknesses in units of microns, and are typical of those found in a modern process. The materials themselves vary with the manufacturer. Various compounds are used at interface layers, and the aluminum is often the material of choice for the

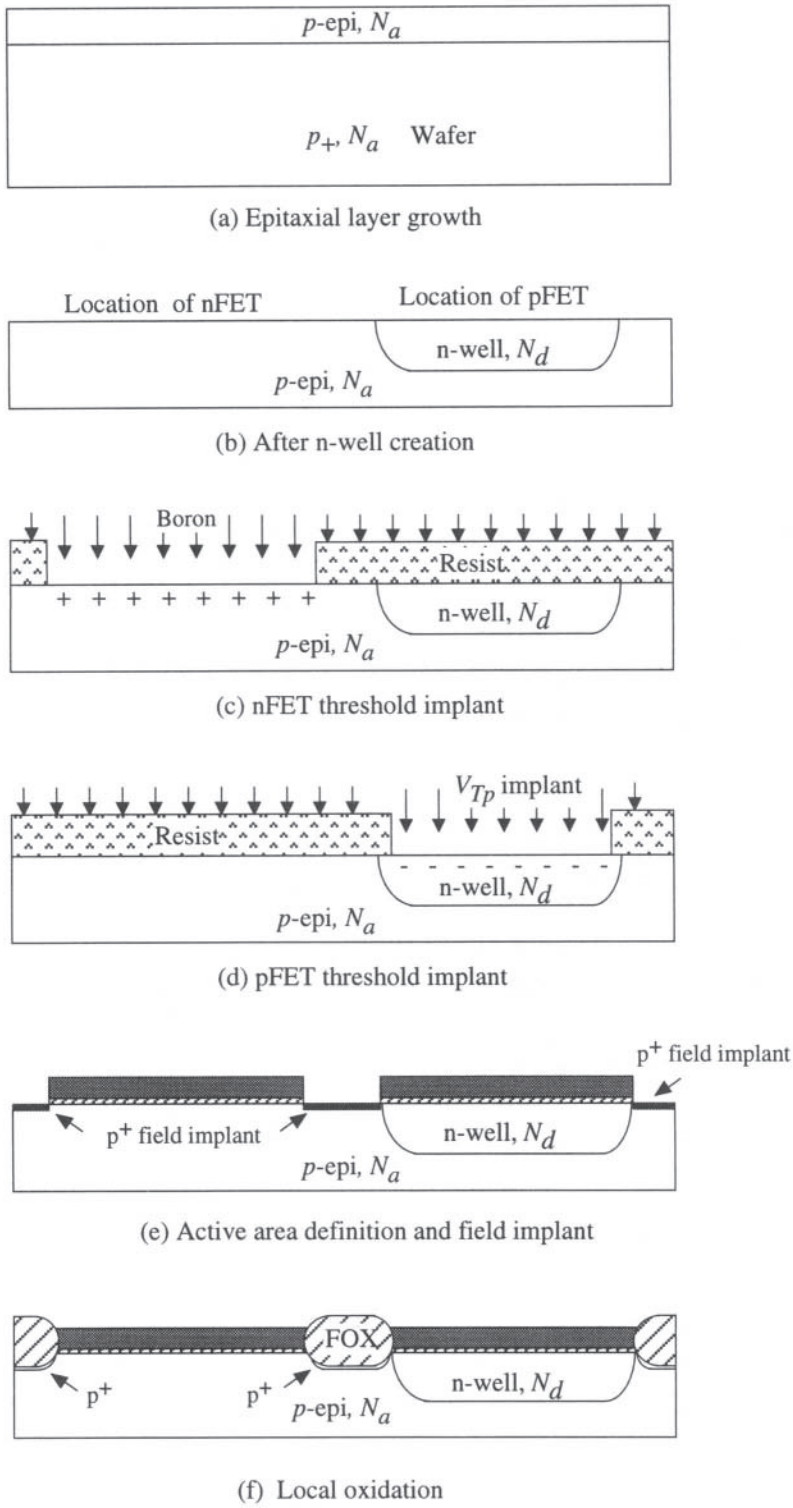


Figure 2.18 Initial steps in a bulk n-well CMOS process

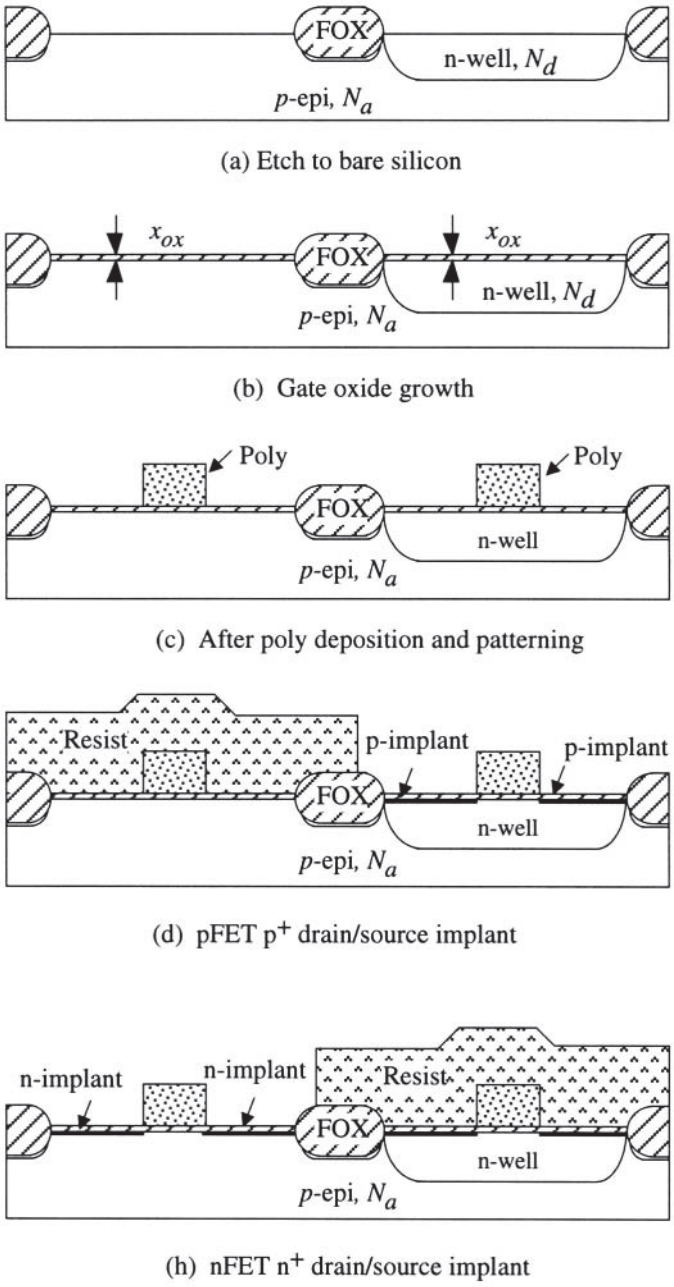
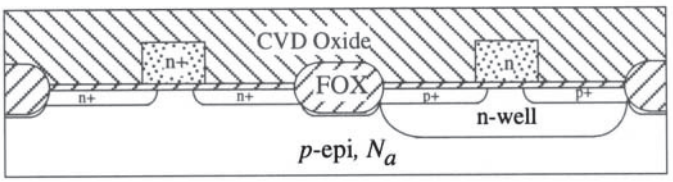
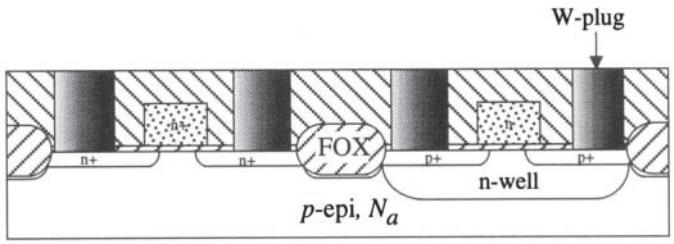


Figure 2.19 Formation of MOSFETs in the CMOS process

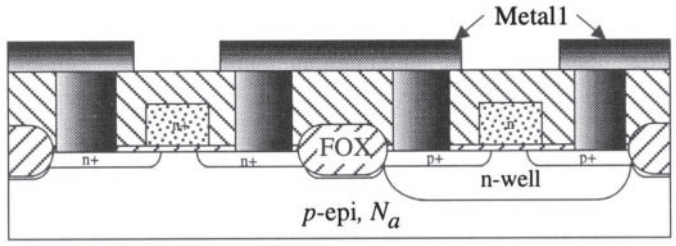
high level layer(s). Although the structures vary, the circuit designer is generally not overly concerned with specifics such as the materials, since it is the electrical characteristics that are of prime importance.



(a) CVD oxide and planarization



(b) Active contacts and tungsten plugs



(c) Metal1 Deposition and patterning

Figure 2.20 First metallization step

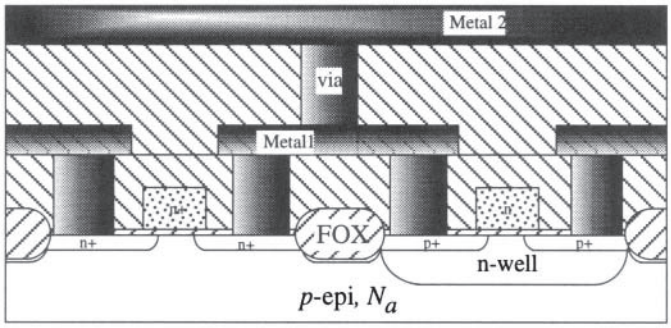


Figure 2.21 Cross-sectional view after second metal deposition

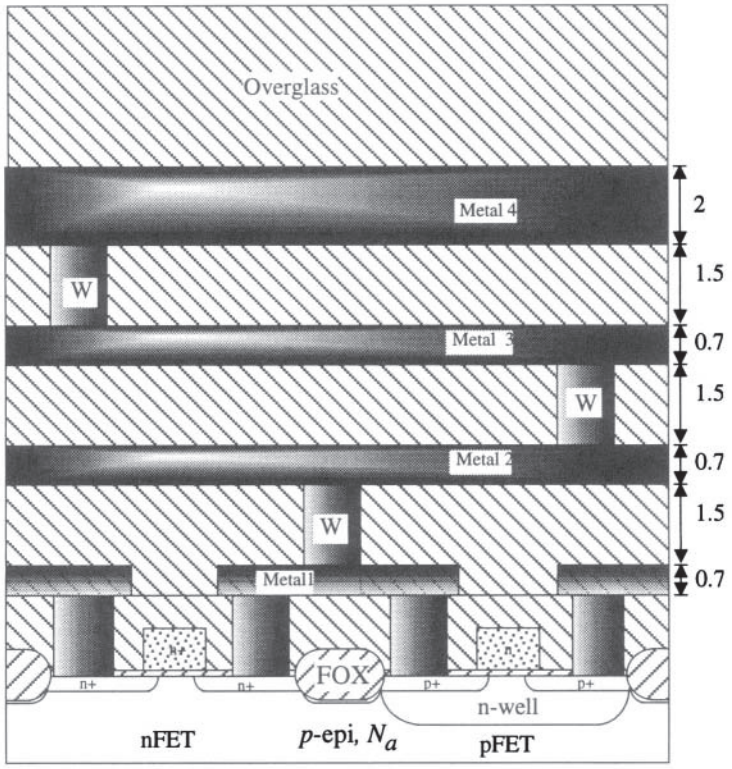


Figure 2.22 Visualization of 4-metal interconnect structure

2.5.1 Silicide Structures

Polysilicon is used for the gate material because the material has excellent coverage, adheres well to the silicon dioxide, and can be doped. Unfortunately, even heavily doped poly has a relatively high resistivity which limits it use as an interconnect. This problem is solved by using a high-temperature (refractory) metal such as titanium as a "coating" on the top, creating what is called a **silicide.**

The sequence in Figure 2.23 shows how a silicide can be created in the basic process flow. After the drain and source implants have been completed, a layer of titanium is patterned on top of the transistors; this yields titanium silicide $TiSi_2$ on both the poly gate and the drain/source regions as shown in Figure 2.23(a). Next, a CVD insulating oxide layer is applied [Figure 2.23(b)]. Contact cuts and tungsten plugs complete the sequence, and results in the structure shown in Figure 2.23(c). In this example, the tungsten plugs are used to contact the silicide drain and source regions. Silicides appear in many different forms and structures, but all have the same objective: reduce the resistance of the polysilicon line.

2.5.2 Other Bulk Technologies

In addition to n-well CMOS, two other bulk processes can be created: p-well and twin-well (also known as twin-tub). These are illustrated in Figure 2.24. The p-well process in 2.24(a) starts with an n-type bulk wafer. pFETs can be placed in the substrate, but a p-well must be added to accom-

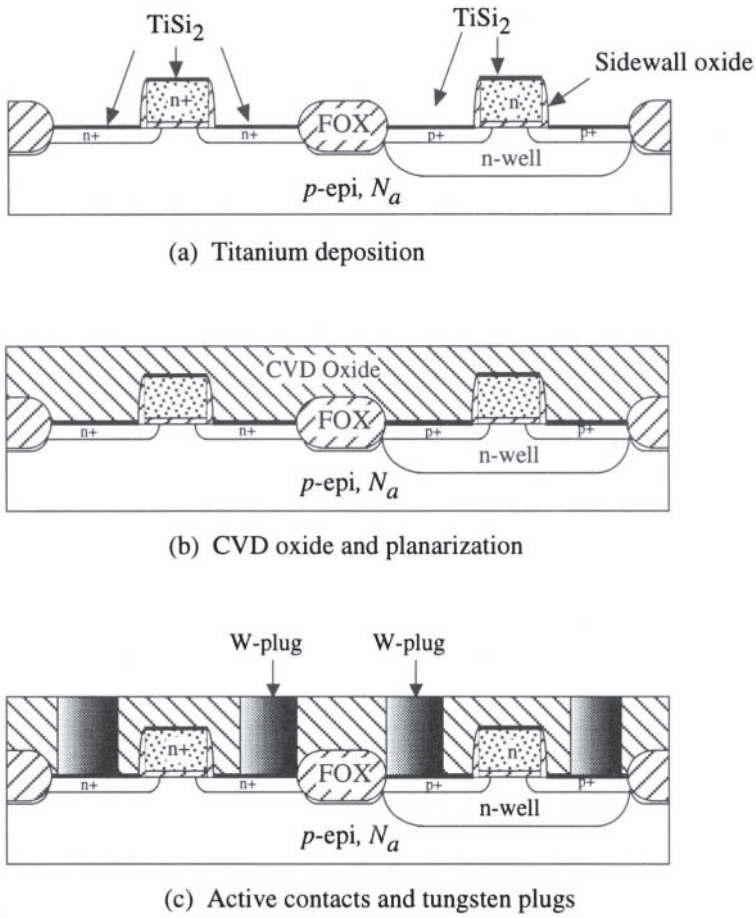


Figure 2.23 Silicides and plug creation in a CMOS process

modate nFETs. In a twin-tub process, a high-resistivity epitaxial layer is grown on top of the wafer, and the separate n-well and p-well regions are introduced for pFETs and nFETs, respectively. This approach is shown in Figure 2.24(b). Another technology called silicon-on-insulator (SOI) has undergone many reincarnations over the past 20+ years. The most recent variation uses an oxide that has been grown on the wafer, with device regions created by selective epitaxial growth above the oxide.

Variations in the technology are done for many reasons: economics, speed, radiation-hardening, and others. For our purposes, we will continue to use the n-well process as being typical in the circuit design process. This provides a good foundation for learning to design in any technology. One must, of course, pay attention to the values of critical parameters whenever switching to a new process. Effects that are negligible in one generation may be critical in the next!

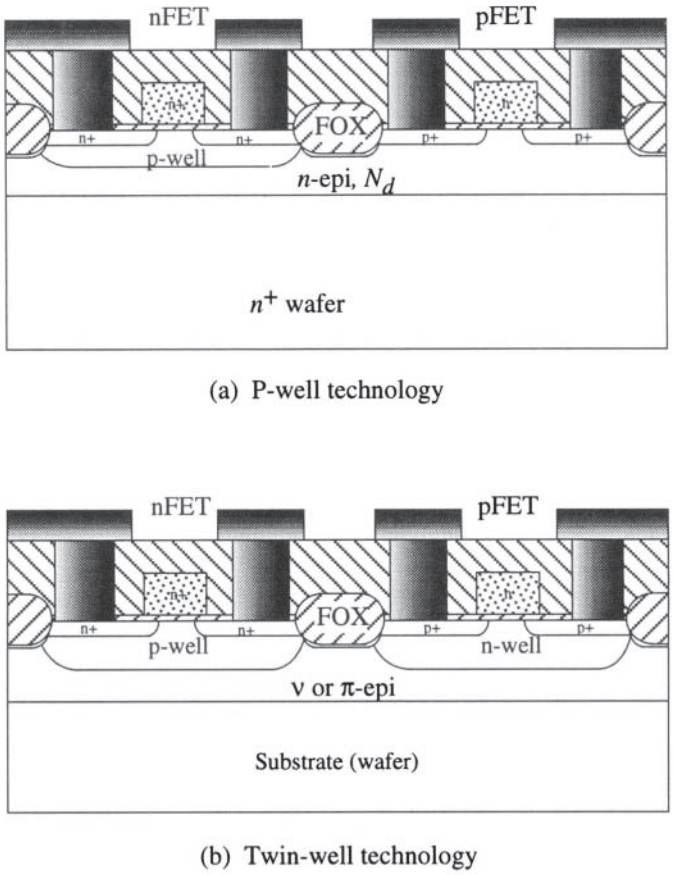


Figure 2.24 Alternate bulk CMOS technologies

2.6 Mask Design and Layout

Chip design centers around two main tasks:

- Translate the necessary logic function into equivalent electronic circuits, and,
- Create fast switching networks.

As we will see in later chapters,[5] synthesizing logic operations is accomplished by proper placement and connection of the MOSFETs, i.e., the circuit topology. Switching performance, on the other hand, is more difficult to control as it depends upon the sizes of the transistors, the characteristics of the circuit connections, and the parasitic resistance and capacitance in the circuit.

Physical design deals with specifying the exact size and location of every geometrical shape on every material layer of the chip. At the design level, this is accomplished by designing every mask that is needed to fabricate the 3-dimensional structure. This is done by using a CAD drawing tool known as a **layout editor** that allows the engineer to specify the pattern of every lithographic step in the process flow. The physical design step gives important characteristics such as the transistor

[5] Circuit and logic design starts in Chapter 3, and is the subject of the remainder of the book.

packing density and the electrical transmission properties of the interconnect "wires" that are possible in a given fabrication process line.

Every patterning step in the process flow requires a separate mask. The masking steps needed in the basic n-well CMOS process that was described above are as follows.

1. **Nwell:** the n-well mask
2. **Active:** regions where FETs will be placed
3. **Poly:** the polysilicon gate pattern
4. **Pselect:** regions where the p-type ion implant will form p^+ regions
5. **Nselect**: regions where the n-type implant will form n^+ regions
6. **Poly contact:** cuts in the oxide that provide Metal1 contacts to the poly
7. **Active contact:** cuts in the oxide that provide connectrions from Metal1 to n^+ or p^+
8. **Metall:** pattern for the first layer of metal
9. Via: oxide cut for Metal1 to Metal2 connections
10. Metal2: pattern for the second layer of metal.

More complicated processes include other layers; for example, additional metal layers are required in advanced circuit designs. However, the general comments here remain valid. It is worthwhile to point out that every conducting layer is separated from the next layer (both above and below) by an insulating oxide. The presence of oxide layers is not denoted explicitly in the mask set listing; they are, however, implied by the contact cut masks such as Poly contact and Via.

The alert reader may have noticed that the mask set listed above does not explicitly list separate masks for the nFET and pFET threshold adjustment implants. This is because the necessary masks are **derived** from others in the group. For example, a mask **Nthresh** used to define the nFET ion implant step can be constructed from the logical expression

$$\textbf{Nthresh} = (\textbf{Active}) \text{ AND } (\textbf{Nselect})$$

as the overlap of the two defines all nFET regions that require the implant. Similarly, the threshold implant pattern for pFETs can be obtained from

$$\textbf{Pthresh} = (\textbf{Active}) \text{ AND } (\textbf{Pselect})$$

using the same reasoning.

In the absence of a layout editor, it is useful to visualize the effect of each mask by means of a set of drawings that show the surface geometry after each step. In Figure 2.25(a), the first mask Nwell defines the locations of the nwells needed for pFETs. The first masking layer is also used to provide **registration marks** (or alignment marks/targets) on the wafer that are used to align several of the masks that follow. The next mask in the sequence defines the active areas as in Figure 2.25(b). The Active mask definition provides flat areas that are used for MOSFETs and substrate or nwell bias contacts. After this is completed, the gate oxide is grown and the polysilicon gate deposition takes place. The poly layer is then patterned using the Poly mask as shown in Figure 2.25(c). In most processes, the poly layer is n-doped *in situ* (meaning that it is doped while being grown) and has a refractory metal layer on top; this allows poly lines to be used as short-run interconnects.

The next masking layers in the sequence are illustrated in Figure 2.26. The drawing in (a) shows the Pselect mask (or Nblock) that defines the silicon regions that are exposed to the boron p-type ion implant. This creates p+ regions in the silicon that are used for both pFETs and low-resistance substrate contacts as shown. The p-implant is followed by the n-type ion implant that has a pattern defined by the Nselect (or Pblock) mask. Nselect gives nFET drain and source regions and n-well contact regions (for applying the power supply voltage). After the FETs are created, the wafer is covered with oxide. The next mask is the Active contact, which is used to define where the oxide cuts are made for connections to the metal1 lines. The addition of these contact cuts is shown in Figure 2.26(c). After the plug-material is deposited into the contact cuts, the first layer of metal is

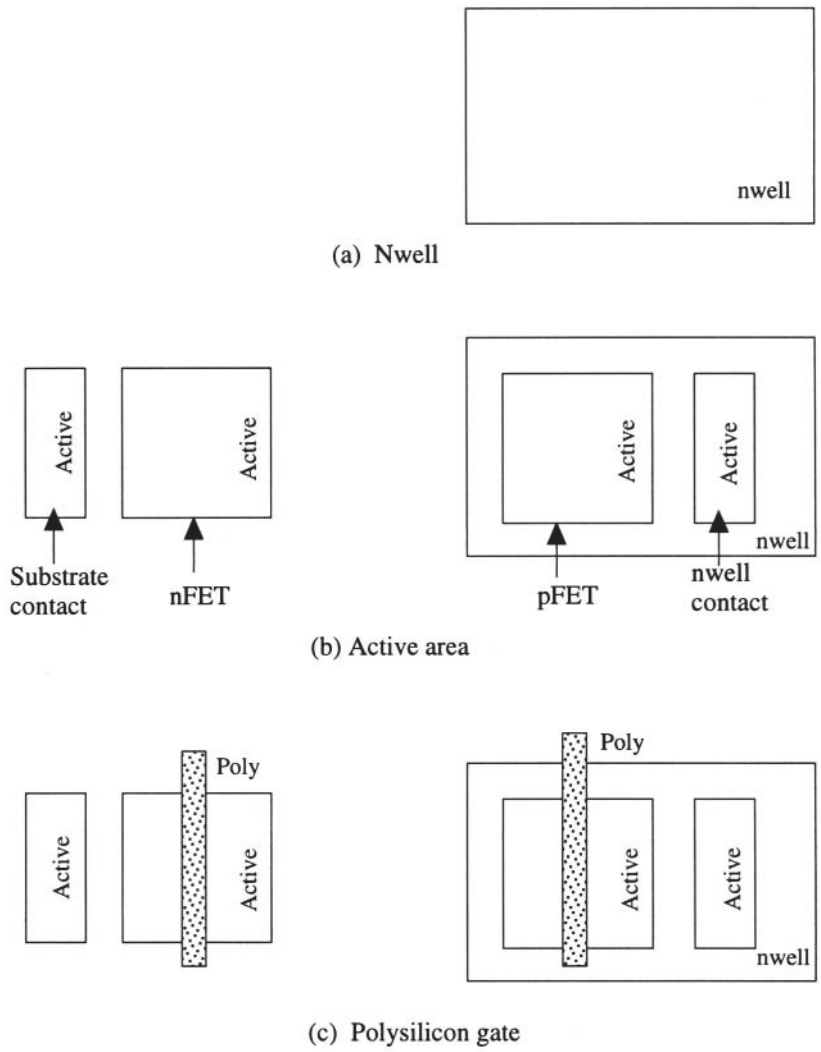


Figure 2.25 Basic masking steps used in defining FETs

deposited over the surface. The final component of the basic mask set is Metal1 which defines the patterns of the Metal1 interconnect layer. This is shown by the drawing in Figure 2.27 where the outlines of the Metal1 features are shown by heavy lines. It should be noted that Metal1 lines are used as an interconnect over the entire chip. The material can be electrically connected to Active areas (n^+ or p^+), polysilicon lines or to higher level metal lines.

The sequence defined by the drawings above only provide examples of the Active contact mask. Figure 2.28 shows the three types of contacts listed in the process flow (Active, Poly, and Via) to more clearly illustrate the characteristics. An Active contact is used as an electrical connection between the drain/source regions of a FET and the Metal1 interconnect layer. A Poly contact is used to connect a Poly line with a Metal1 line. Finally, a Via gives an electrical connection between Metal1 and Metal2 lines. The drawing also illustrates that Metal1 and Metal2 can overlap without shorting. The same statement holds for Poly to both Metal1 and Metal2.

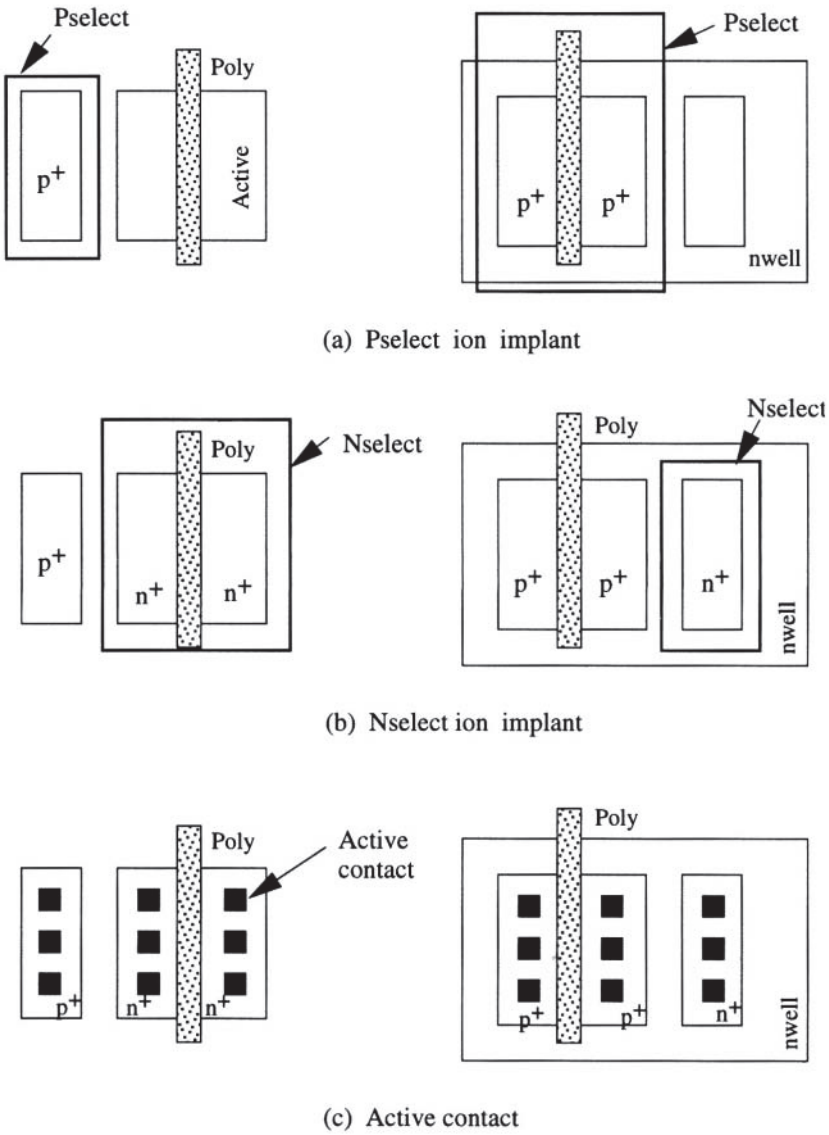


Figure 2.26 Remaining masking steps used to define FETs

2.6.1 MOSFET Dimensions

The values of the channel width W and channel length L combine to give the aspect ratio (W/L), which is (as we will see) the primary design parameter in CMOS integrated circuits. The process flow above allows us to see the relationship between the masks and the physical dimensions of the final device. These are shown in Figure 2.29. Consider first the channel length L. As discussed previously, we may write

$$L = L' - 2L_o \tag{2.29}$$

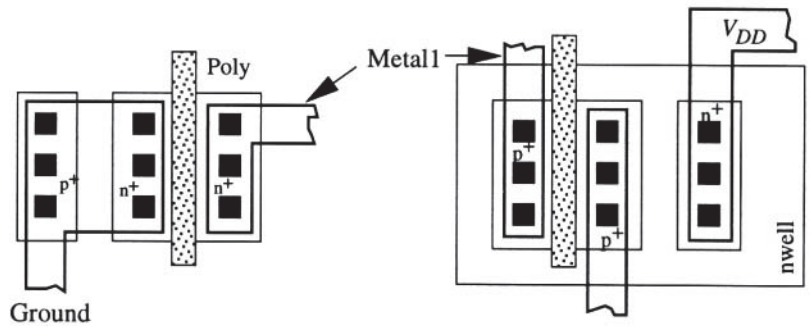


Figure 2.27 Metal1 patterning

where L' is the drawn channel length and L_o is the overlap due to lateral doping effects. In general, the drawn channel length L' is established by the minimum linewidth of the polysilicon gate, while L_o is a result of the processing recipe. We therefore conclude that the resolution of the polysilicon gate mask determines the minimum channel length of a transistor. Often one refers to the characteristics of a particular process by referring to the value of L, the electrical channel length. For example, a "0.35 micron process" usually implies that L=0.35μm, from which one might extrapolate that the poly linewidth is around $L' \approx 0.40 \mu m$ for a simple FET process. Alternately, a "0.35 micron process" might specify the lithographic resolution for the gate, i.e., the drawn channel length L'.

The channel width W is set by the dimensions of the Active mask, since this defines the width of the region that will accept the Nselect or Pselect implants into the silicon. The complicating factor in LOCOS isolation is that encroachment decreases the size of the active area from the original Active mask definition. Defining the encroachment amount as ΔW, then channel width may be expressed as

$$W = W' - 2(\Delta W) \tag{2.30}$$

where W' *is* the drawn width of the Active mask. The aspect ratio of a transistor is thus given in terms of the drawn dimensions by

$$\left(\frac{W}{L}\right) = \frac{W' - 2(\Delta W)}{L' - 2L_o} \tag{2.31}$$

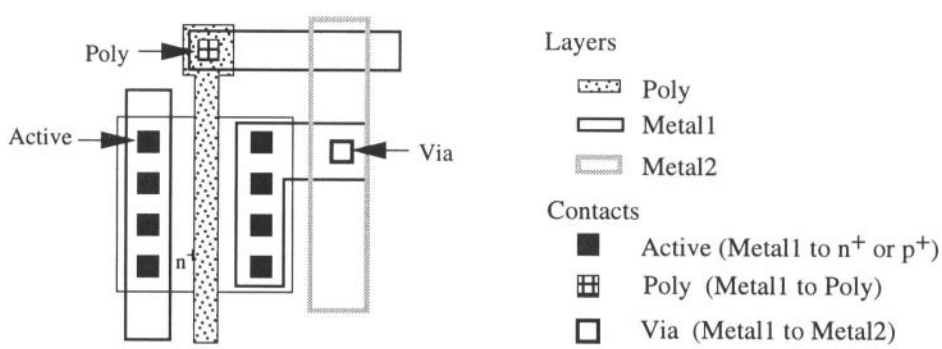


Figure 2.28 Contacts, vias, and interconnect layers

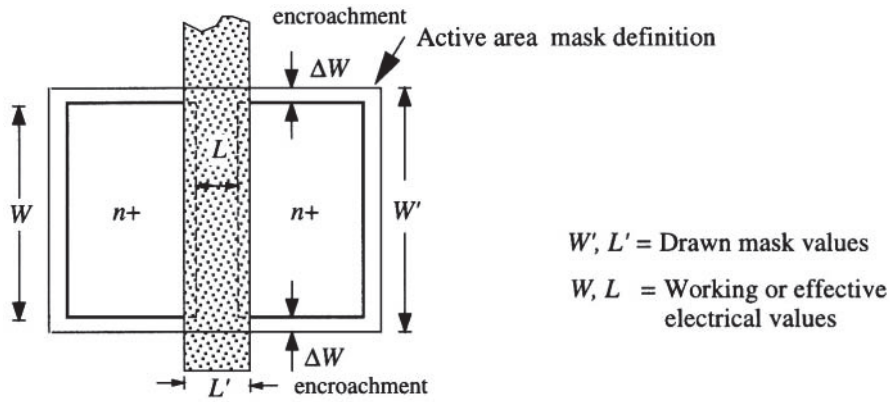


Figure 2.29 Geometrical definitions for a MOSFET

It is this value that must be used in the device equations, as it represents the electrical dimensions seen by the current flow lines.

In practice, one often uses the drawn values *W'* and *L'* as extracted from the layout as the basic input parameters into a circuit simulation program. The electrical values are then calculated in the simulation code using addition input from the processing values. This allows one to work directly with the layout without having to visualize the differences between drawn and effective (electrical) values. For example, a SPICE simulation may use a statement such as

```
Minput 20 5 0 0 NFET W=10U L=0.5U
```

where $10\mu m$ and $0.5\mu m$ are the *drawn* values. The electrical values used in the program are calculated from data listed in the model statement

```
.MODEL NFET <parameters>
```

in the parameter listing. The model itself defines the relationship between the drawn and effective values. SPICE-related literature often refers to the effective values L_{eff} and W_{eff}, which are in fact the true electrical lengths that are defined by the device geometry, not the layout masks. An example can be seen in the .MODEL example listing provided at the end of Chapter 1.

In this book, we will continue to use the notation *W* and *L* to denote the electrical values, and *W'* and *L'* as the drawn values. This convention for *W* and *L* is consistent with that used in device physics, and also keeps our equations simple by not having to add so many subscripts!

2.6.2 Design Rules

Design rules are a listing of critical geometrical size and spacing constraints that must be observed when designing a lithographic mask pattern. Each rule originates from limitations imposed by items such as the lithographic process, equipment characteristics, and physical considerations. Every process flow is characterized by a distinct set of design rules that are derived from the relevant limitations that arise in the manufacturing equipment, physical properties of the materials, or critical circuit parameters. Failure to adhere to every rule can mean that the mask set will not result in a functional chip.

2.6.3 Types of Design Rules

Although a design rule set may be quite long and involved, most rules can be grouped into a few major classes. These are summarized below.

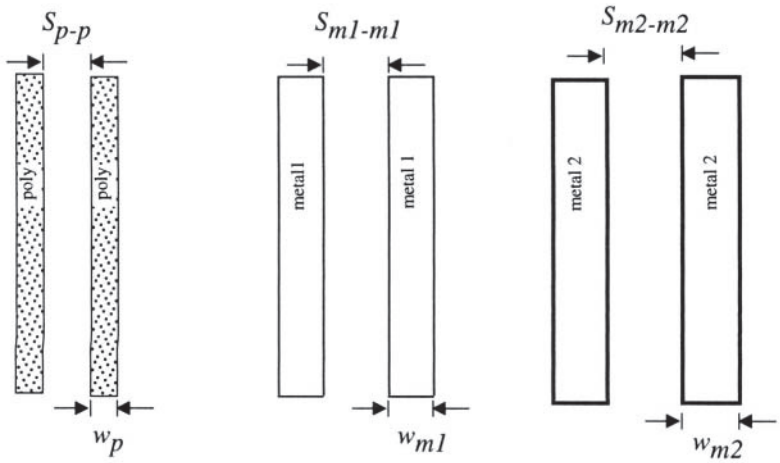


Figure 2.30 Linewidth and spacing rules

Minimum Width and Spacing

The first group deals with the minimum width for defining a line on a layer, and the minimum spacing to an adjacent line on the same layer. Figure 2.30 shows these values for lines created on layers of polysilicon, metal1, and metal 2. Minimum width specifications are shown as w_p, w_{m1}, and w_{m2} respectively. Minimum spacing distances are denoted by $S_{p\text{-}p}$, $S_{m1\text{-}m1}$, and $S_{m2\text{-}m2}$ in the same order. Similar specifications apply to every masking layer in the process, including nwell, active, nselect, pselect, and so on.

Exact Size and Surround

An exact size rule dictates the dimensions of a particular object on the mask. In CMOS processing, exact size rules arise in specifying the dimensions of oxide cuts that are used to provide access between two conducting regions. These are usually square, or close to square, and are due to considerations in the equipment characteristics. An example is shown in Figure 2.31(a) where the contact is specified to have dimensions x_c by y_c. The spacing $S_{c\text{-}c}$ is also indicated in the drawing.

A surround rule is required when a feature is to be placed in a region that has already been pat-

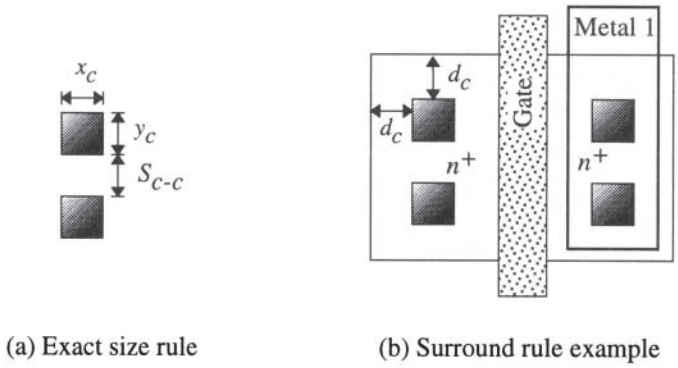


Figure 2.31 Exact size and surround rule examples

terned by a previous masking step. Consider the MOSFET shown in Figure 2.31(b). The shape of the n^+ region is determined by the ACTIVE and NSELECT masks; the active contact must reside within the borders of the region. The surround spacing d_c is used to compensate for small registration errors in the alignment during the exposure and insure a working contact.

MOSFET Rules

Self-aligned MOSFETs require an additional set of rules to insure that the devices will operate. These are generally required to compensate for any misalignment between a mask and the features already on the die.

Gate overhang is shown in Figure 2.32 as D_{GO}, and shows the distance that the gate must extend beyond the Active area. Recall that the self-aligned MOSFET uses the gate as a mask to the n^+ or p^+ ion implant. The overhang distance insures that the doped regions formed by the implant are physically separated even if the poly gate mask is not perfectly aligned to the existing active area region. An extension distance D_E must be used when the active border changes as seen in the upper left side of the device. This is included for the same reason as D_{GO}, namely, to allow for registration errors (in the horizontal direction) of the poly gate mask.

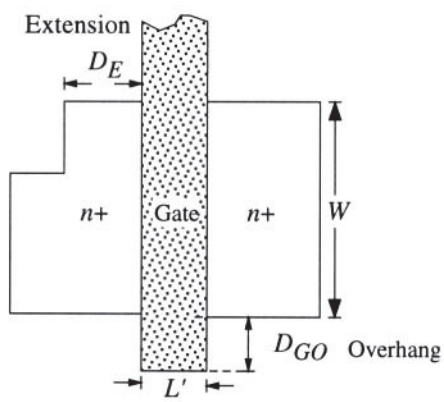


Figure 2.32 Two basic MOSFET design rules

Active Contacts

One important point that arises from the exact size specification of active contacts is the use of multiple contacts between metal and n^+ or p^+ silicon regions. This is particularly important in MOSFETs for the reduction of parasitic resistance and the proper operation of the device.

Consider the FET layout shown in Figure 2.33(a). Both the source and drain regions use multiple active contacts with the design rules specifying the size and spacing of the contacts. The cross-sectional view along the line X-X' in Figure 2.33(b) shows the details of the connections. One reason for using many contacts is to reduce the effective **contact resistance** between the metal and the semiconductor. Each contact point is characterized by a resistance R_c. If we use m-contacts, then the individual contributions are all in parallel as shown in Figure 2.33(c). This reduces the effective value to

$$R_{c,\,eff} = \frac{R_c}{m} \tag{2. 32}$$

which will in turn help the circuit to switch faster.

The second reason that multiple contacts are used is to insure that the current flow between the

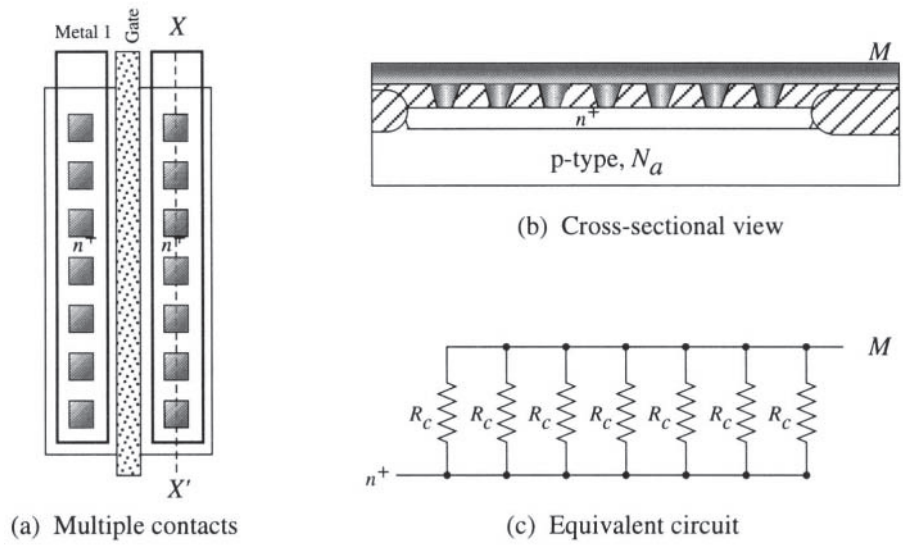


Figure 2.33 Use of multiple FET contacts

drain and source is spread out over the entire electrical width *W* of the FET. This is portrayed in the two drawings of Figure 2.34. In figure (a), only one contact is used for each side of the device. From basic electrostatics, the current flow is along electric field lines that must satisfy all boundary conditions. Since the field lines originate and terminate at the contacts, the flow is concentrated in the region directly between the two contacts. This implies that the current flow density is lower as we move away from the location of the contacts. In terms of MOSFETs, this says that the electrical channel width will be smaller than the geometrical channel width *W*. This, of course, is undesirable as it negates the design implied by the layout. The use of multiple contacts as shown in drawing (b) overcomes this problem by spreading out the current flow lines over the entire extent of the device. These arguments illustrate that multiple contacts are necessary whenever the design rule set specifies exact sizes for oxide cuts that access active regions.

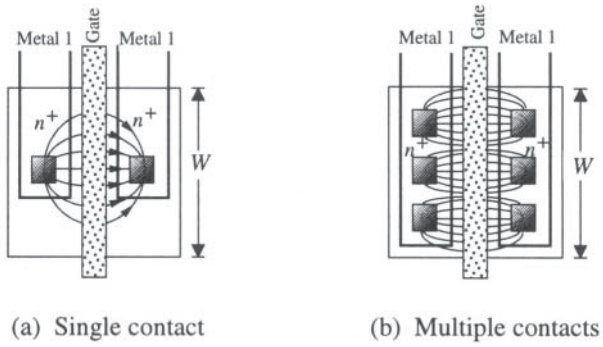


Figure 2.34 Current flow paths between contacts

2.6.4 General Comments

Design rule listings are process-specific. Numerical values for every required mask dimension are derived by considering details such as the capabilities of the lithographic imaging equipment, and physical effects such as pn junction depletion widths and coupling parasitics. Since they are concerned with minimum sizes, the design rule set acts as the limiting factor for circuit integration. A critical observation is the fact that the size of a MOSFET is almost negligible compared to the area consumed by interconnect lines and other wiring. The means that the area of a CMOS layout tends to be limited by the interconnect itself.

CMOS design is directly related to layout, but in modern circuit engineering, the need to master layout varies with the scope of the position. Some chip designers "push polygons" on a daily basis, while others let layout technicians provide amazingly compact solutions that can be analyzed and simulated. The philosophy as to whether a chip designer should or should not perform the layout seems to vary with the company, or even depend upon the "culture" of a particular group within a company. If you are a student and just learning the subject, it is generally accepted that "the more you know, the better off you are."[6]

2.7 Latch-Up

The structure of a bulk CMOS process introduces a problem known as latch-up in which the circuits fail to operate and the chip draws excessive power supply current. In practice, this may arise in two different situations:

- The chip is operating normally, and then goes into the latch-up state. The only way to restore normal conditions is to disconnect the power supply, and then reapply. The chip may go into latch-up again.
- The chip goes into latch-up immediately upon application of the power supply.

In the worst case scenario, the chip will be destroyed by heat. In the early days of CMOS development, latch-up was a major problem that slowed the growth of the technology. Although the factors that induce the condition are now understood, there are times when it can still be a problem.

The origins of latch-up can be shown using the drawing in Figure 2.35(a). This identifies four distinct layers and the current flow path from the power supply voltage (V_{DD}) to ground associated with the latch-up condition. Under normal operating conditions, current along this path would only consist of leakage components. However, the nature of the layering scheme gives rise to parasitic bipolar transistors as shown in Figure 2.35(b). As discussed below, the bipolar transistors form a feedback loop that may induce latch-up.

The left drawing in Figure 2.36 shows how the layers 1 through 4 in the CMOS structure can be viewed as creating a 4-layer device with the pnpn layering. In power electronics, this is called a **silicon-controlled rectifier** (SCR) where it is used as a switching device between the top (p) and bottom (n) regions. From the qualitative viewpoint, the 4-layer structure blocks current flow from the power supply to ground due to the presence of reverse-biased pn junctions in the path. However, if one of the internal regions (n-well or p-epi) can be electrically shorted, then we would be left with a forward biased pn junction from the top to the bottom. This would give I_{DD} exponential characteristics of the form

$$I_{SS} \approx I_S \exp\left(\frac{V_{DD}}{kT/q}\right) \tag{2.33}$$

indicating a large current flow. To describe this type of behavior, we will construct an equivalent

[6] So you should sign up for the next VLSI systems course that is offered!

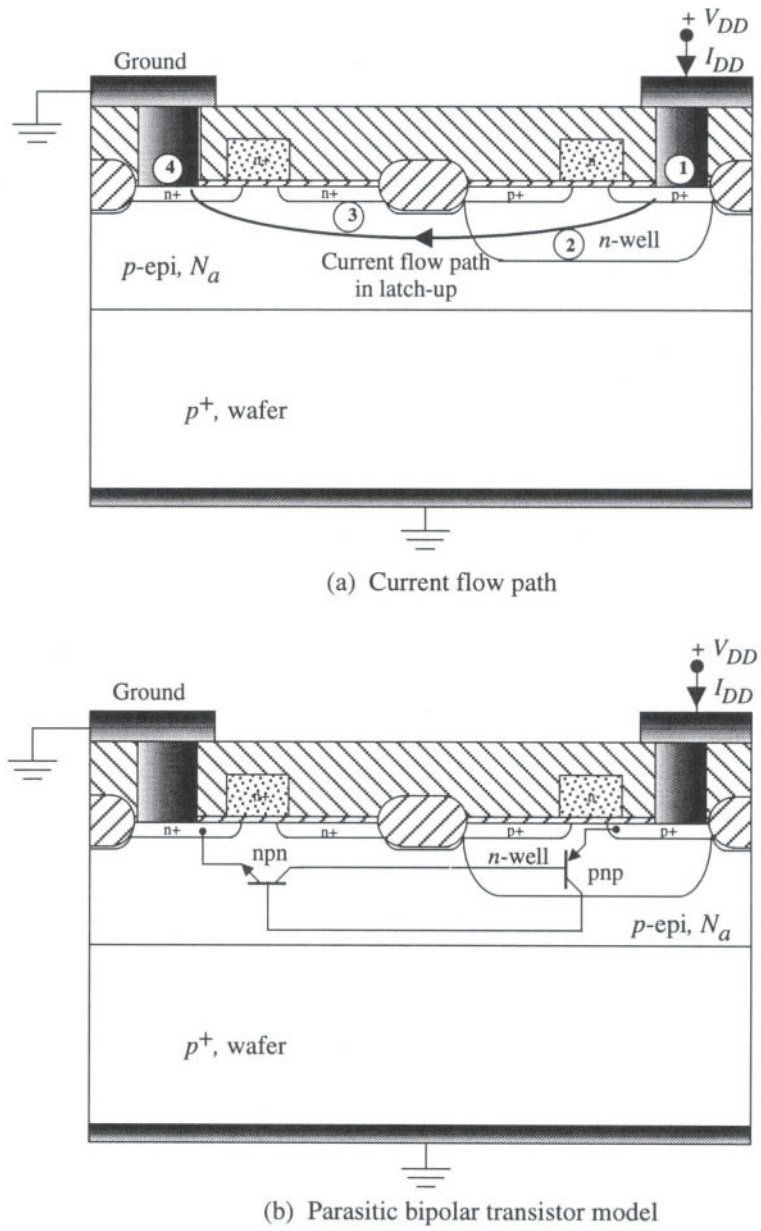


Figure 2.35 Latch-up current flow paths in the n-well structure

pair of npn and pnp bipolar transistors from the structure as shown on the right side of Figure 2.36. This shows that the base of the pnp also acts as the collector of the npn transistor. Similarly, the collector of the pnp is electrically the same as the base of the npn. The pnp-npn pair thus form a set of coupled bipolar transistors where the current flow through one affects the conduction characteristics of the other.

The effect of the parasitic bipolar transistors can be understood using the equivalent circuit shown in Figure 2.37(a). Resistors have been added to represent the parasitic effects of the silicon regions. This circuit diagram illustrates the fact that the two transistors are connected in a manner that creates a feedback loop between the them; this is the middle window in the circuit schematic.

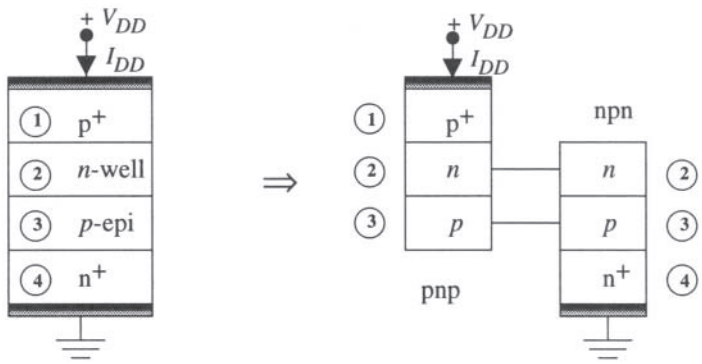


Figure 2.36 4-layer modelling of the latchup network

Consider the current I_1 on the right-hand side of the circuit. Under normal conditions, this will be leakage current. However, it creates a voltage V_1 that acts as an emitter-base voltage $V_1=V_{EB1}$ across the pnp transistor, enhancing the flow of I_{Ep} and I_{Cp} in the device. Most of this current contributes to I_2 which establishes the voltage $V_2=V_{BE2}$ across the base-emitter junction of the npn transistor. This in turn enhances the collector current I_{Cn}, which increases I_1 and completes the feedback loop. If both transistors are conducting, then the structure may induce the latchup condition.

A plot of I_{DD} as a function of V_{DD} is shown in Figure 2.37(b). For small values of V_{DD}, the current is restricted to small leakage levels and there is no problem with the structure. However, if V_{DD} is increased to a value V_{BO}, known as the **break-over voltage,** the blocking characteristics of the 4-

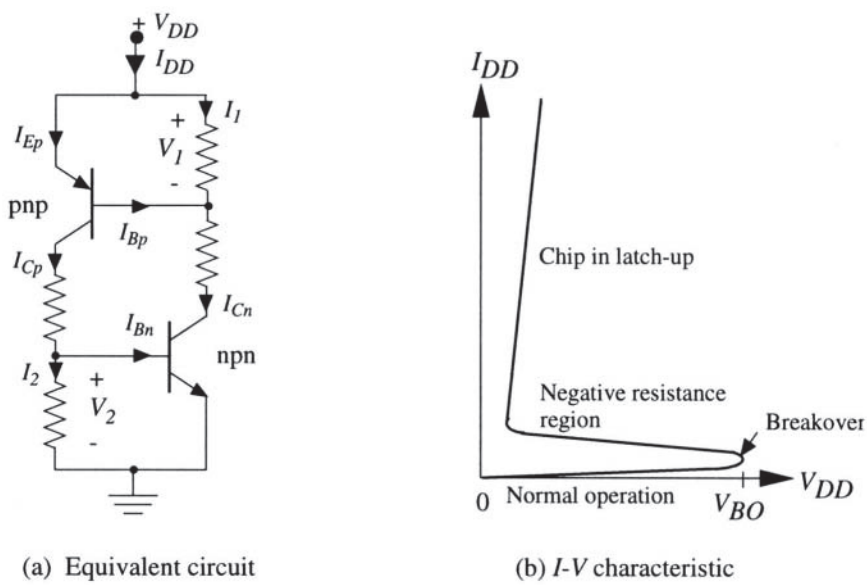


Figure 2.37 Bipolar transistor modelling of latch-up

layered pnpn device break down, allowing I_{DD} to increase as the voltage V_{DD} drops. Since this has a negative slope, it is called the **negative conductance region** of the *I-V* curve. The voltage falls until the device "catches" the exponential curve, which describes the high current flow levels from the power supply to ground. The chip is classified as being in latch-up in this region.

An analysis of the circuit shows that the critical condition for latch up is when the sum of the common base current gains is equal to 1, i.e.,

$$\alpha_{npn} + \alpha_{pnp} = 1 \tag{2. 34}$$

where

$$\alpha_{npn} = \frac{I_{Cn}}{I_{En}} \quad , \qquad \alpha_{pnp} = \frac{I_{Cp}}{I_{Ep}} \tag{2. 35}$$

Recall that the forward alpha α_F of a bipolar transistor depends on the emitter current as illustrated by Figure 2.38. For small values of I_E, α_F increases as the forward injection builds. It then levels out for a range of currents, and finally exhibits a roll-off at high currents due to high-level injection effects. For the present situation, we note that the current levels are initially restricted to small leakage currents, which in turn gives small α_F values. However, since the current flow is enhanced by the feedback mechanism, both α_{npn} and α_{pnp} increase. This illustrates how the common-base current gain condition in equation (2.34) can be reached: the feedback loop enhances both the npn and the pnp current, which in turn increases the gain.

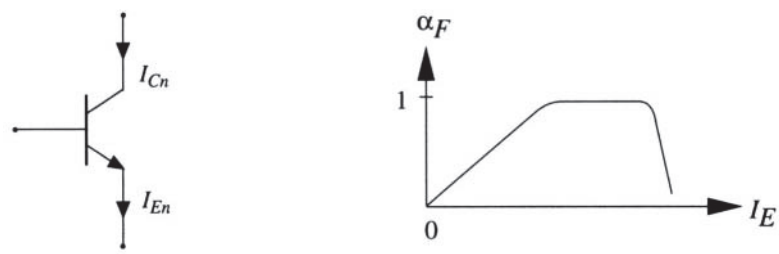


Figure 2.38 Common-base current gain in a bipolar junction transistor

2.7.1 Latch up Prevention

Latch up prevention is accomplished by designing the structures in a manner that acts against the formation of the feedback network. Various techniques have been developed and are usually included in the design rule specifications.

Consider, for example, the cross-sectional view in Figure 2.39 where we have added resistors to represent the path seen by the current flowing through the semiconductive regions. One way to break the feedback loop between the npn and the pnp transistors is to insure that the effective resistance r_x is very large. This can be achieved by using trench isolation (instead of LOCOS) to provide a glass block between the two parasitic devices. Another effective deterrent is to insure that the substrate resistance r_{sub} is very small, since this would effectively break the path between the collector of the npn and the base of the npn transistors. This implies that the wafer doping should be heavily p-type, as in our earlier example of a CMOS process flow.

Other rules to prevent latchup can be summarized by the following statements that apply to an

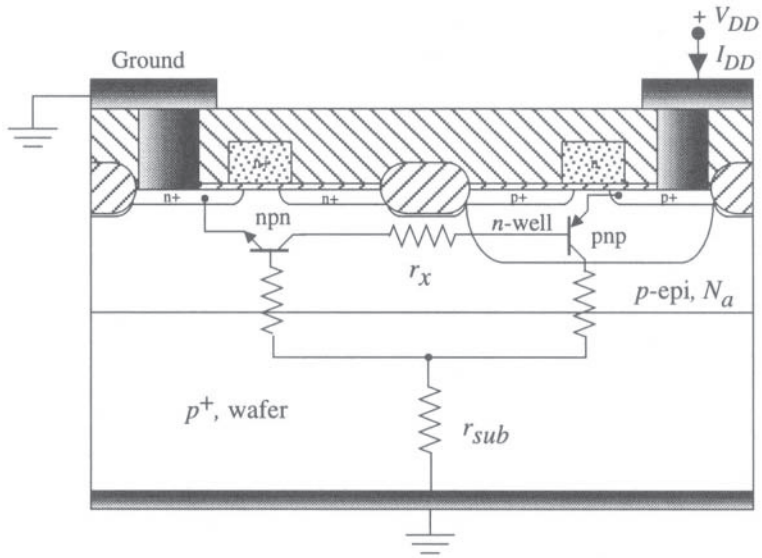


Figure 2.39 Semiconductor parasitics in the latch-up paths

n-well process. Each is portrayed in the various aspects of Figure 2.40.

- Apply ground-to-substrate contacts whenever possible;
- Add V_{DD}-to-n well contacts whenever possible.

These help eliminate up the voltage drops that might bias the bipolar transistors into the active operational region.

- Obey all design rule spacings, especially those that affect the formation of the parasitic BJTs.

This one is aimed at reducing the gain of the parasitic bipolar transistors by making the BJT base

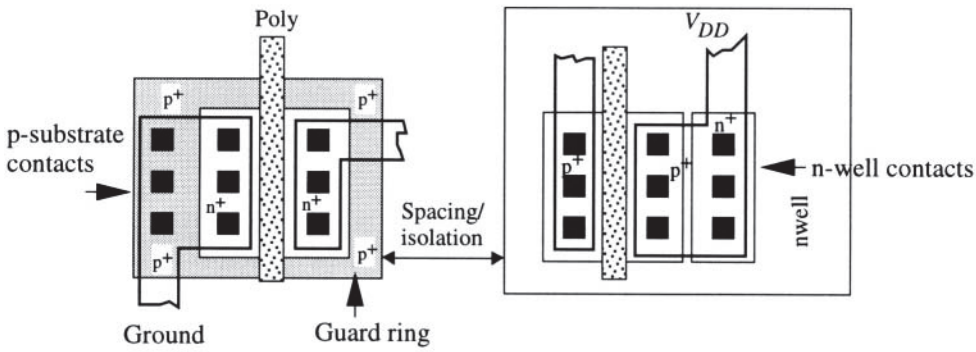


Figure 2.40 Layout for latchup prevention

widths large.

- Use **guard rings** around devices or groups of same-polarity devices.

A guard ring is a doped region that surrounds the MOSFET(s) and is biased by the power supply (if it is an n-type ring) or ground (for the case of a p-type ring). The physical extent of the guard ring increases the BJT base widths, while the bias helps maintain well-defined potentials. Rings help avoid latchup, but do consume chip real estate.

2.8 Defects and Yield Considerations

High-density chip designs consist of a few tens of millions of MOSFETs. This has become so commonplace that the technical achievement of silicon processing is generally overlooked. Consider the implications of having a "good die". This means that, so far as the testing process has shown, every circuit in the die operates as it should. In other words, every important feature of every transistor and the interconnect wiring has the correct behavior and, therefore, the correct structure.

In the real world of semiconductor manufacturing, we are continually faced with the fact that only a percentage of the die are functional. This is expressed by the **yield** Y of a process such that

$$Y = \frac{\text{Number of Good Die}}{\text{Number of Total Die}} \times 100\% \tag{2.36}$$

Obviously, a high yield is required to have a profitable design. The study of **yield enhancement** centers on the problem of achieving this goal. While most of the problems originate in the fabrication process and are the responsibility of those involved in the process flow definition and design, some factors have a direct effect on the chip and circuit designer. An example is in the formulation of a design rule set, since these are derived to insure functional chips that can be manufactured within the limits of the equipment. The second most important concept is that of the area A of an individual die.

Silicon wafers cannot be manufactured without random defects on the surface. These are usually specified using the defect density D that has informal units of # of defects per cm^2. In a highly simplified analysis, we may state that the presence of a defect within a die boundary will lead to a non-functional circuit. In order to maintain a reasonable yield, this implies that the die area A should be kept as small as possible. The concept can be expressed by the simple equation

$$Y \sim e^{-\sqrt{DA}} \times 100\ \% \tag{2.37}$$

which is based on an empirical model and is drawn in Figure 2.41. The reasoning for this dependence can be understood using the drawings in Figure 2.42. Suppose that a given wafer is used for two chip designs with the area A_1 of design 1 larger than the area A_2 of design 2: $A_1 > A_2$. In Figure

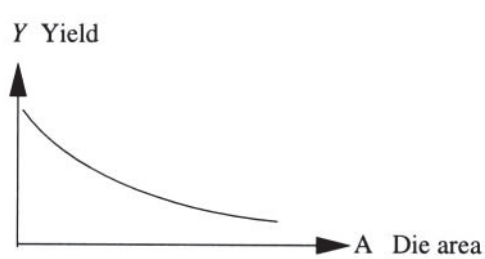


Figure 2.41 Yield as a function of die area

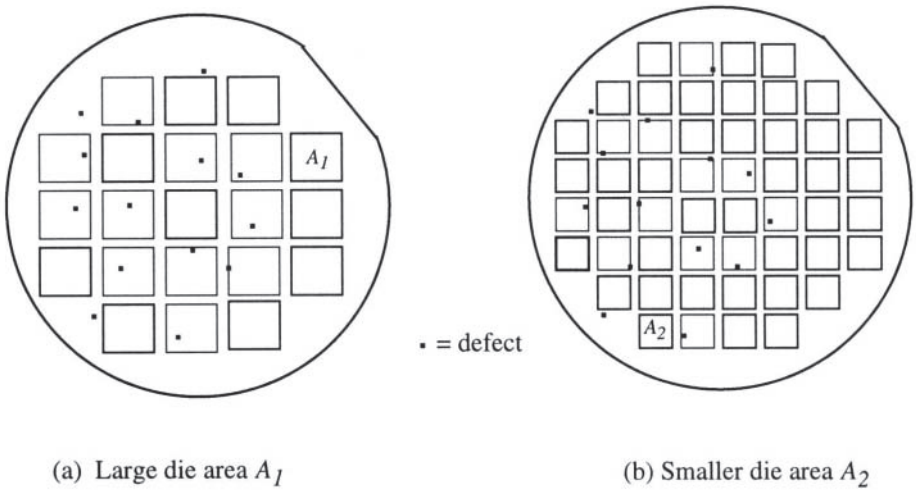


Figure 2.42 Dependence of die size on yield

2.42(a), the large die area implies that there is a high probability of overlaying a defect, which reduces the yield. However, the smaller area design in Figure 2.42(b) reduces the probability that a die boundary will surround a defect, thus increasing the yield.

This problem highlights one of the main aspects of modern CMOS VLSI designs. To maximize the yield, we want to use a small die area. This in turn requires that we have

- the ability to achieve small lithographic features,
- compact layout of the circuits, and
- efficient algorithms that allow us to compact the maximum amount of logic into as small an area as possible.

To the chip designer, the latter two problems are of paramount importance. The physical design is the manifestation of a logic network, but every device, wire, and connection requires surface area. Even though two circuits may give the same logical output, the internal characteristics dictate how large the circuit will be and the factors that limit the response. This is one of the main themes we will follow throughout this book.

2.8.1 Other Failure Modes

Several failure modes exist in MOS integrated circuits. These originate in the fabrication sequence, and cannot be eliminated completely. Short-term (immediate) problems include line breaks and metallization failures, lithographic problems, and short circuits. Long-term effects are more difficult to characterize and may require an extensive analysis of the problem. Gate oxide shorts are unique to MOSFETs, and deserve a more detailed explanation.

Gate Oxide Shorts

A gate oxide short (GOS) nullifies the field effect, and thus renders a MOSFET non-functional. Qualitatively, a GOS is due to the failure of the oxide to act as an insulator between the gate and the substrate. The origin of this type of defect is shown in Figure 2.43. During the initial stages of the growth of the gate oxide, it is possible that a defect or surface non-uniformity inhibits the local growth rate of the SiO_2 layer as shown in (a); this yields a pinhole in the oxide. Depositing the poly over a pinhole yields the MOSFET shown in (b). Assuming that the poly gate is doped n-type, then

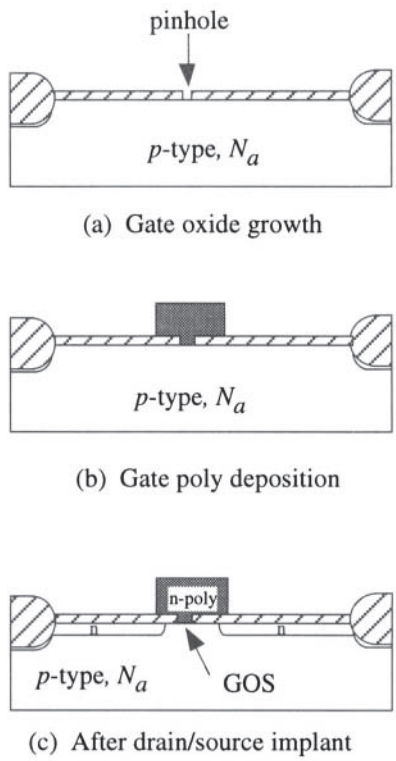


Figure 2.43 Gate-oxide shorts in MOSFETs

the GOS is electrically equivalent to a pn junction, and can be modelled using an equivalent circuit where the gate and substrate are connected by a reverse-biased diode. The prohibits the formation of the drain-source channel inversion layer, and the FET will not function as intended.

Consider the formation of the oxide itself. If the oxidation process is allowed to continue (yielding a thick oxide), then it is highly probable that the pinhole will "fill up" and not be a problem. However, the probability of a having GOS increases as thinner oxides are used. With modern devices having gate oxide thicknesses x_{ox} less than 70-80 $\AA$, this type of failure mechanism becomes increasingly important. In static logic circuits (discussed in the next three chapters), it is possible to use a special testing technique known as "I_{DDQ}-Approach" which is particularly adept at finding clusters of GOS failures. The interested reader is directed to the literature for more information on this subject.

2.9 Chapter Summary

High-speed CMOS circuit design is intimately related to the structures that can be fabricated on a silicon substrate and then be transferred to a manufacturing line for mass production. Many of the limitations found in modern chip design are related to the fabrication process.

As we have seen in our short discussion, the electrical characteristics of FETs are established by a fairly complex interplay of parameters and dependences in the process flow. Experienced designers always examine how these values affect the performance of a logic network, and try to work with a given set of electrical characteristics to achieve their goals. We will adhere to this philosophy through the remaining chapters of this book.

2.10 References

The circuit designer usually views the fabrication technology as the basis for layout. There are many excellent books on the subject. A few select titles are listed below, but additional literature can easily be found in a library or using the resources of the world-wide web.

[1] S. Campbell, **The Science and Engineering of Microelectronic Fabrication,** Oxford University Press, New York, 1996.

[2] C.Y. Chang and S.M. Sze, **ULSI Technology,** McGraw-Hill Book Company, New York, 1996.

[3] B. Ciciani (ed.), **Manufacturing Yield Evaluation of Vlsi/Wsi Systems,** IEEE Computer Society, 1995.

[4] J-P. Colinge, **Silicon-on-Insulator Technology,** Kluwer Academic Publishers, Boston, 1990.

[5] D. De Cogan, **Design and Technology of Integrated Circuits,** John Wiley & Sons, New York, 1990.

[6] S.K. Ghandhi, **VLSI Fabrication Principles,** 2nd ed., John Wiley & Sons, New York, 1994.

[7] R. K. Gulati and C.F. Hawkins (eds.), **IDDQ Testing of VLSI Circuits,** Kluwer Academic Publishers, Boston, 1993.

[8]

[9] N. J. Jha and S. Kundu, **Testing and Reliable Design of CMOS Circuits,** Kluwer Academic Publishers, Boston, 1990.

[10] M. Madou, **Fundamentals of Microfabrication,** CRC Press, 1997.

[11] R. Rajsuman, **Iddq Testing for Cmos Vlsi** , The Artech House, 1995.

[12] M. Sarrafzadeh and C.K. Wong, **An Introduction to VLSI Physical Design,** McGraw-Hill Book Company, New York, 1996.

[13] N. Sherwani, **Algorithms for VLSI Physical Design Automation,** Kluwer Academic Publishers, Norwell, MA, 1993.

[14] S.M. Sze, **VLSI Technology,** 2nd ed., McGraw-Hill Book Company, New York, 1988

[15] R. Troutman, **Latchup in CMOS Technology,** Kluwer Academic Publishers, Boston, 1986.

[16] J. P. Uyemura, **Physical Design of CMOS Integrated Circuits Using L-Edit,** PWS Publishers, Boston, 1995.

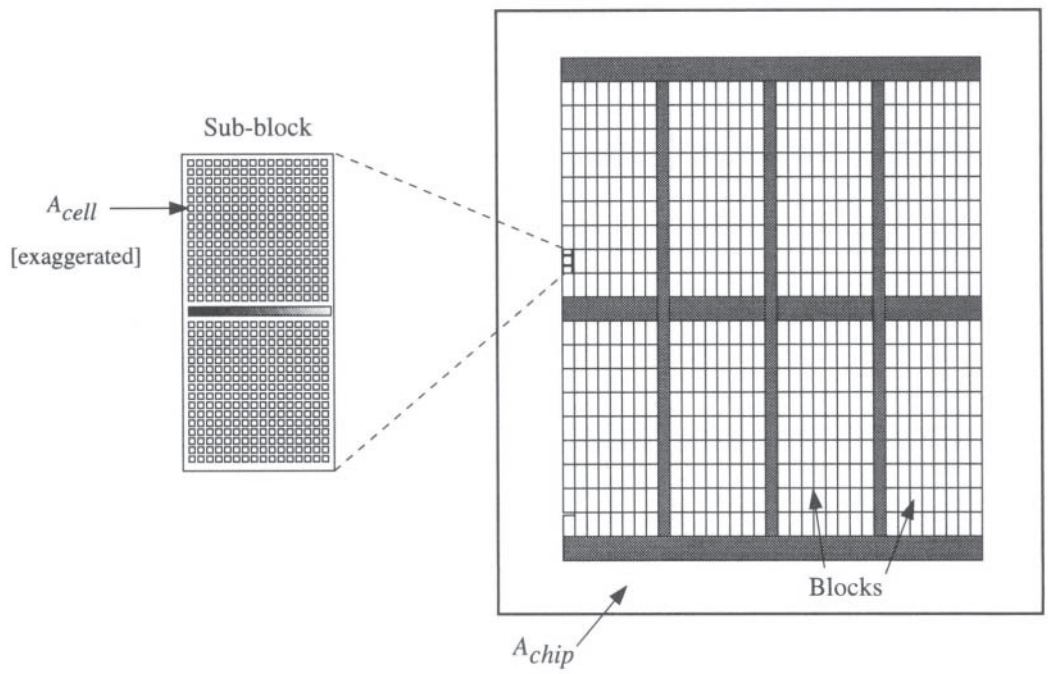


Figure 7.26 Chip layout for a DRAM array

- increase the permittivity ε of the dielectric;
- decrease the distance of separation d between the plates; or ,
- increase the surface area A_s of the plates.

Modern DRAM cell design employs all three techniques along with some additional considerations to increase C_s to a reasonably large value.

Composite Insulators

Classically, silicon dioxide (generically called **oxide** here) has served as the insulating dielectric in MOS capacitors. This is due to many reasons, including ease of growth, excellent insulating characteristics, and uniform coverage. However, the permittivity of silicon dioxide is relatively small

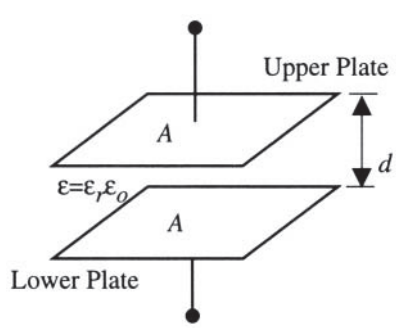


Figure 7.27 Parallel-plate capacitor

($\varepsilon_{ox} \approx 3.9\varepsilon_o$), and extremely thin oxides are subject to electrical breakdown[7] and tunnelling problems, so that there has always been interest in finding an insulator that has superior characteristics. One popular material is silicon nitride (Si_3N_4 or simply **nitride**) which has a permittivity of $\varepsilon_N \approx 7.8\varepsilon_o$. Nitride has been well studied because of its extensive use in device isolation techniques, and as a passivation layer that covers and protects the finished die. If we replace an oxide insulator with a nitride layer, then the capacitance approximately doubles. There are, however, overriding technology problems that arise.

One approach to using this technology is the ON (oxy-nitride) structure shown in Figure 7.28. The plates of DRAM storage capacitors are made from polysilicon, which is easy to oxidize using thermal growth techniques. Insulating nitride layers are more difficult to create, but can be grown using rapid thermal nitridation (RTN) where a nitrogen gas flow is established over the poly surface and heat is used as a catalyst for the reaction. Oxide is then deposited over the nitride layer which helps to "seal" the layer by filling pinholes. The top polysilicon plate competes the structure. The main idea is to create a composite insulator with a thickness

$$d_{ON} = d_{ox} + d_{Nitride} \tag{7. 118}$$

that has an effective permittivity

$$\varepsilon_{ON} = x\,\varepsilon_{ox} + y\,\varepsilon_{Nitride} \tag{7. 119}$$

where x and y are multipliers that are determined by the relative thicknesses. The actual value of the permittivity will be in the range

$$\varepsilon_{ox} < \varepsilon_{ON} < \varepsilon_{Nitride} \tag{7. 120}$$

so that the structure will exhibit a larger capacitance than if only a simple oxide insulator of the same thickness were used. ONO and other structures have also been published in the literature. While these techniques help increase the value of C_s, they are not sufficient by themselves to boost the capacity to the necessary values.

Storage Capacitor Design

Modern storage cell design techniques center around increasing the capacitor plate area by creating 3-dimensional capacitor structures. This leads to larger values of C_s without increasing the foot-

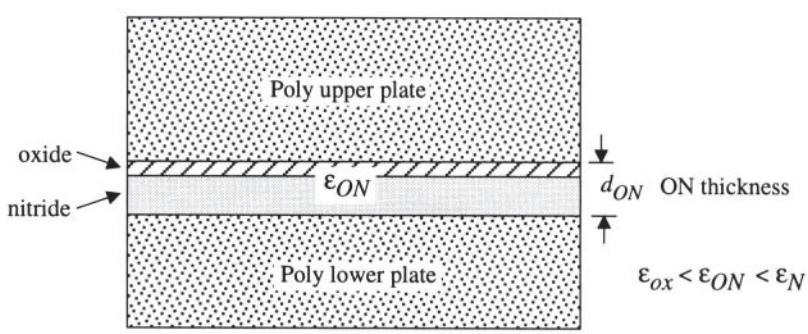


Figure 7.28 Structure of an ON-insulator capacitor

[7] Oxide breakdown is discussed in Chapter 10 in the context of input protection networks.

print. There are two main approaches. One is to create a **trench capacitor** by first etching away a portion of the of the silicon substrate, and then constructing the capacitor using the walls of the trench. The other is to build the capacitor above the substrate plane, creating what is called a **stacked capacitor.** Although both are found in practice, stacked capacitors are more common than trench structures.

An integrated pair of DRAM cells that use trench capacitors is illustrated in Figure 7.29. Let us examine Cell 1 on the left side in more detail to understand the structure. The lower plate of the storage capacitor is the extended n^+ region implanted in the silicon substrate; this becomes one side of the access transistor MA1, which is connected to the bit line on the other side. The storage cell itself is created by etching a trench into the silicon substrate and then oxidizing the surface to give the insulating oxide layer. Doped polysilicon is deposited as a filler, and also acts as the upper capacitor plate. A metal layer then provides electrical contact to the top plate of the cell.

The philosophy behind the trench capacitor structure is easily understood by referring to the simplified geometry illustrated in Figure 7.30. It is seen that the plate area of the capacitor is given by

$$A_s = A_{bot} + A_{side} \tag{7.121}$$

This shows explicitly that the plate area is larger than the footprint area

$$A_{footprint} = A_{bot} \tag{7.122}$$

because of the sidewall area A_{side}. With the dimensions shown in the drawing, $A_{bot}=XY$ while

$$A_{side} = 2XD + 2YD \tag{7.123}$$

where D is the depth of the trench. Deep trenches are useful for increasing the capacitance, but are more difficult to fabricate.

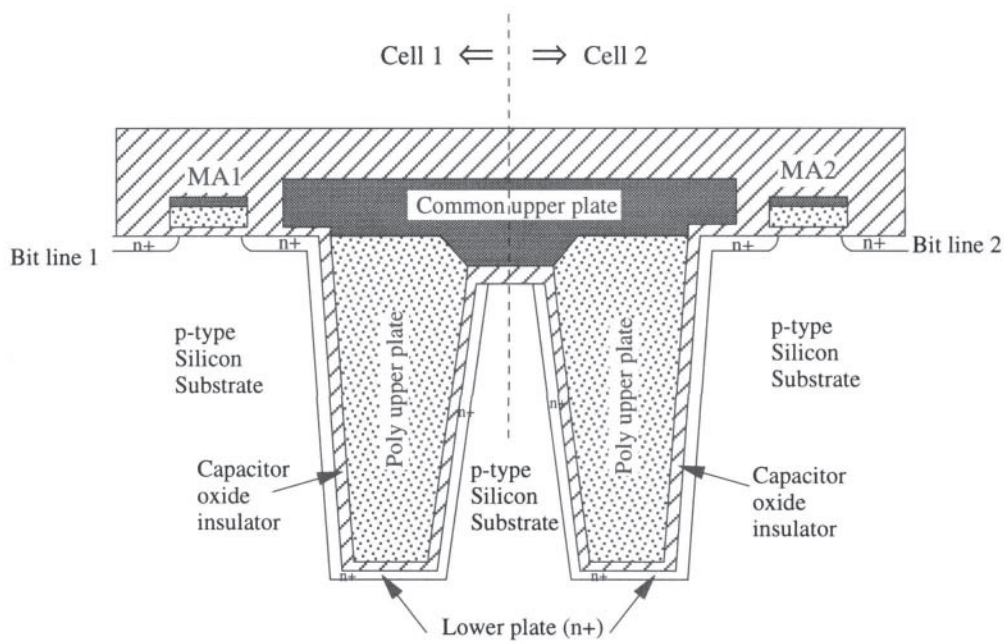


Figure 7.29 Integrated DRAM cells with trench capacitors

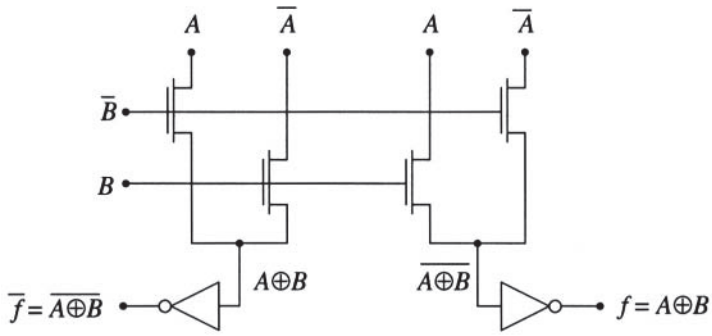


Figure 9.34 CPL XOR/XNOR 2-input array

using the networks shown in Figure 9.34.

It is possible to create other functions using the 2-input array as a basis, but it is important to remember that dual-rail logic requires that both the output f and its complement $\overline{f}$ need to be formed. Consider the array in Figure 9.35. By applying the logic rules, the left side evaluates to

$$f_1 = \overline{A} \cdot B + \overline{B} \cdot \overline{C} \tag{9.29}$$

while the right side gives

$$f_2 = A \cdot B + \overline{B} \cdot C \tag{9.30}$$

We see that f_2 is in fact the complement of f_1; this observation can be verified by using a simple truth table listing. However, caution must be exercised when designing logic functions as the placement of the input variables becomes critical. In a more general case, two outputs g_1 and g_2 should be checked to insure that they are in fact complements of one another. If not, we must either generate $\overline{g}_1$ and $\overline{g}_2$, using separate circuitry, or include the stages in a logic cascade where the final result at the end of the chain produces complementary outputs.

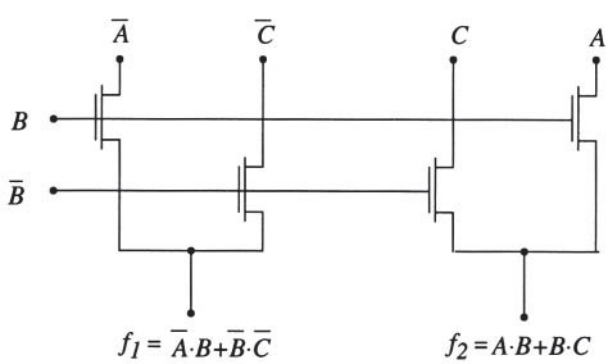


Figure 9.35 General logic arrangement in CPL

Layout

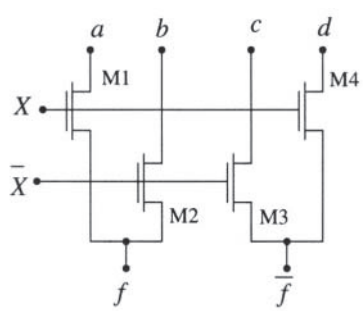
One striking feature of CPL arrays is the simplicity of the layout. One layout strategy for 2-input arrays is shown in Figure 9.36. This approach uses horizontal-oriented FETs and allows simple linear gates for the control variables X and $\overline{X}$. The metal routing is arbitrary at this point. If one is designing a library cell, then the input and output port locations should carefully selected so as to allow simple cascades and wiring. CPL has the advantage that the 2-input array layout can be used for any function pair AND/NAND, OR/NOR, or XOR/XNOR by routing the input signals to the proper nodes (a, *b, c, d,* and X, $\overline{X}$). In other words, the circuit topology is invariant, and the signal placement determines the actual logic function that the circuit performs. This aspect is even more intriguing when we note that we can optimize the circuit for switching time, and then maintain many aspects of the speed for every function.

9.4.2 3-Input Arrays

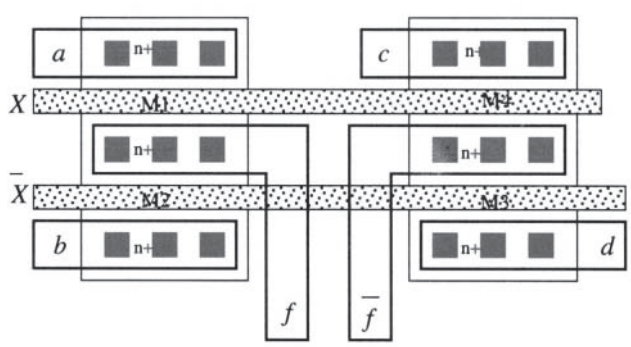
CPL also provides for 3-input gates using a structured approach. Figure 9.37 shows the switching network for a 3-input AND/NAND array. To understand the logic construction, consider first the left array. Applying the rules gives

$$A \cdot B \cdot C + A \cdot \overline{A} + B \cdot \overline{B} = A \cdot B \cdot C \tag{9.31}$$

directly. The right hand array evaluates to



(a) Circuit diagram



(b) Layout

Figure 9.36 2-input array layout example

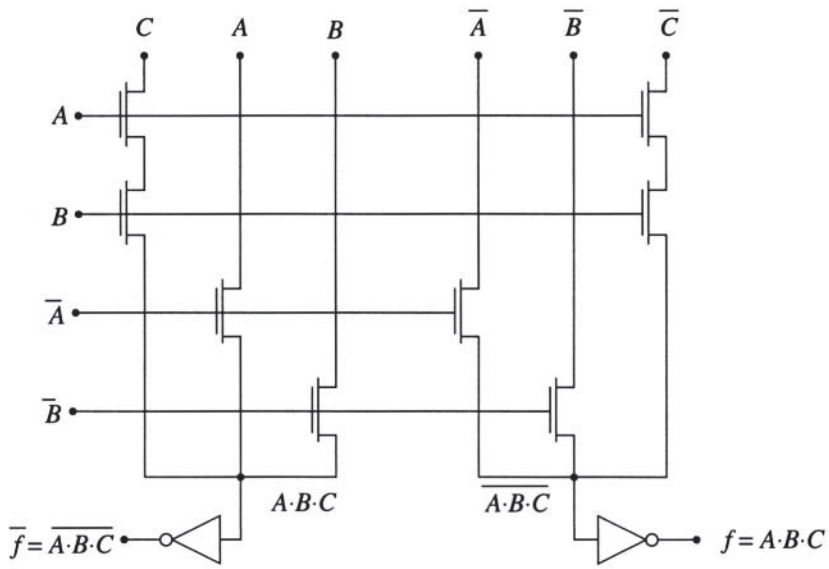


Figure 9.37 3-input AND/NAND array in CPL

$$\begin{aligned} \overline{A} \cdot \overline{A} + \overline{B} \cdot \overline{B} + A \cdot B \cdot \overline{C} &= \overline{A} + \overline{B} + A \cdot B \cdot \overline{C} \\ &= \overline{A} + \overline{B} + \overline{C} \\ &= \overline{A \cdot B \cdot C} \end{aligned} \tag{9. 32}$$

which is the NAND3 operation. We thus see that the AND/NAND pair can be built using only 8 nFETs to provide the logic. Inverters may be added at the output to restore the logic 1 voltage and speed up the circuit response.

A 3-input OR/NOR array can be obtained by simply interchanging the variables applied to the gates, which results in the circuit shown in Figure 9.38. The left side provides the OR function as verified by the reduction

$$\begin{aligned} \overline{A} \cdot \overline{B} \cdot C + A \cdot A + B \cdot B &= \overline{A} \cdot \overline{B} \cdot C + A + B \\ &= A + B + C \end{aligned} \tag{9. 33}$$

Similarly, the right hand array gives

$$\begin{aligned} A \cdot \overline{A} + B \cdot \overline{B} + \overline{A} \cdot \overline{B} \cdot \overline{C} &= \overline{A} \cdot \overline{B} \cdot \overline{C} \\ &= \overline{A + B + C} \end{aligned} \tag{9. 34}$$

and represents a NOR3 logic gate. It is worth mentioning again that the OR/NOR array uses the same transistors arrangement as the AND/NAND. The only difference between the two arrays is in the order of the input variables, just as in the case of the 2-input arrays.

One important aspect of the 3-input arrays is that the longest signal path on either side requires the transmission through 2 series-connected nFETs. This leads to the circuit shown in Figure 9.39.