Fundamentals of MOS Digital Integrated Circuits

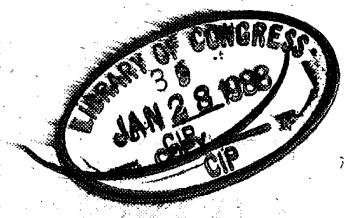
JOHN P. UYEMURA
GEORGIA INSTITUTE OF TECHNOLOGY

ADDISON-WESLEY PUBLISHING COMPANY
Reading, Massachusetts • Menlo Park, California • New York
Don Mills, Ontario • Wokingham, England • Amsterdam • Bonn
Sydney • Singapore • Tokyo • Madrid • Bogotá
Santiago • San Juan

This book is in the **Addison-Wesley Series in Electrical and Computer Engineering**

Library of Congress Cataloging-in-Publication Data
Uyemura, John P. (John Paul), 1952–
Fundamentals of MOS digital integrated circuits.

Includes index.
1. Digital integrated circuits. 2. Metal oxide semiconductors. I. Title
TK7874.U94 1988 621.381'73 87-11458
ISBN 0-201-13318-0



ABCDEFGHIJ-HA-898

Contents

CHAPTER 5

MOS Integrated Circuit Fabrication

A VLSI processing facility constitutes the major financial commitment for integrated circuit production. Vast amounts of money and engineering time are invested in an effort to advance and refine the state of the art. Higher throughput, circuit densities, and yields are all important to the economic stability of the industry.

Circuit design is structured by the processing. The interplay between chip fabrication and circuit design cannot be separated. Rather, they merge to become a single discipline. Process engineers must understand circuit design, and circuit designers must have a knowledge of fabrication if they are to be effective.

The fundamentals of silicon chip fabrication are discussed in this chapter, with an emphasis on MOS-based processes, as suggested by the chapter title. The level is introductory and is not meant to replace a detailed course in the subject. Instead, the objective is to bring out important points relevant to circuit design.

Persons with experience in the subject matter may choose to simply glance through most of the chapter in a first reading. The mathematics can generally

be omitted without loss of continuity in later chapters. Note, however, that some of the drawings are worth attention.

5.1 Overview of Basic Silicon Processing Steps

This section will center around a brief examination of the basic wafer processing steps used to fabricate silicon integrated circuits. The discussion is intended only as an overview. A reader desiring a more in-depth treatment should consult one of the texts referenced at the end of the chapter.

5.1.1 Thermal Oxidation

Thermal oxidation is the process used to create a "native" silicon dioxide (SiO_2) layer on a silicon surface. The SiO_2 grown by this process is used in different ways. For example, it serves as the insulator in the MOS system, is needed to perform patterning during the lithography, and allows for the technique of LOCOS (local oxidation of silicon) device isolation.

A silicon dioxide layer may be grown by passing an oxygen-rich gas over a silicon surface. Pure oxygen can be used as described by the reaction

$$Si + O_2 \rightarrow SiO_2.$$

Steam is also used to grow oxide layers. The reaction is

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

where it is seen that free hydrogen is left as a by-product. Both reactions use heat as a catalyst, thus giving rise to the term "thermal" oxidation. Oxide growth using O_2 is called a "dry" process to distinguish it from "wet" oxidation in steam.

The kinetics of the oxide growth process can be understood by referring to Fig. 5.1. Figure 5.1(a) shows that the oxygen can react directly with the surface

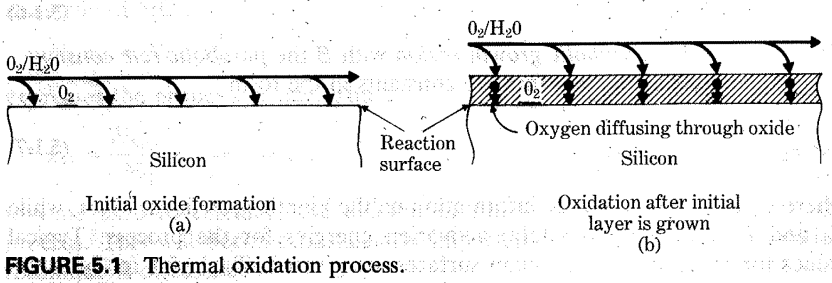


FIGURE 5.1 Thermal oxidation process.

silicon atoms to form SiO_2. This gives a high initial oxidation rate. After an oxide layer forms, the oxygen must *diffuse* through the existing SiO_2 (as shown in Fig. 5.1b) before it can react with the silicon atoms. Consequently, the oxide growth rate decreases from its initial value. An analysis of the reactant fluxes using a simple steady-state model for the growth kinetics shows that the oxide thickness x_{ox} as a function of time t is approximated by the quadratic equation

$$x_{ox}^2(t) + Ax_{ox}(t) = B(t + t_0), \tag{5.1-1}$$

where A and B are quantities characterizing the growth variables, such as temperature and pressure, and t_0 is included to account for any existing oxide at time $t = 0$. Denoting this initial oxide thickness by $x_{ox}(0)$, t_0 is computed from

$$t_0 = \frac{1}{B}[x_{ox}^2(0) + Ax_{ox}(0)]. \tag{5.1-2}$$

The solution to the oxide growth equation is

$$x_{ox}(t) = \frac{A}{2}\left[\sqrt{1 + \frac{4B}{A^2}(t + t_0)} - 1\right], \tag{5.1-3}$$

which allows for the growth rate (dx_{ox}/dt) to be calculated.

To study the oxidation process, assume that $x_{ox}(0) = 0$ so that $t_0 = 0$ for simplicity. Then eqn. (5.1-3) reduces to

$$x_{ox}(t) = \frac{A}{2}\left[\sqrt{1 + \frac{4Bt}{A^2}} - 1\right]. \tag{5.1-4}$$

For small values of t corresponding to the initial oxidation, the square-root term may be expanded in a Taylor series to give

$$x_{ox}(t) \simeq \frac{B}{A}t. \tag{5.1-5}$$

Since the oxide thickness is proportional to the oxidation time t, this is called the *linear growth region* with (B/A) the *linear rate constant.* As the oxidation time is increased, $(4Bt/A^2)$ dominates the square-root term, so for large values of t,

$$x_{ox}(t) \simeq \sqrt{Bt}. \tag{5.1-6}$$

This describes the *parabolic growth region* with B the *parabolic rate constant.*

It is possible to express the rate constants in the form

$$\begin{aligned} B &= C_1 e^{-E_1/kT}, \\ (B/A) &= C_2 e^{-E_2/kT}, \end{aligned} \tag{5.1-7}$$

where C_1 and C_2 contain the information on the kinetic growth variables, while E_1 and E_2 are (approximately) activation energies for the process. Typical values for (111) and (100) silicon surfaces are given in Table 5.1 for both dry

TABLE 5.1 Oxide Growth Rate Constants*

Parabolic Rate Constant: $B = C_1 e^{-E_1/kT}$
Linear Rate Constant: $(B/A) = C_2 e^{-E_2/kT}$

Source	(111) Silicon	(100) Silicon
Dry O_2	$C_1 = 7.72 \times 10^2$ [μm^2/hr]	
	$C_2 = 6.23 \times 10^6$ [μm/hr]	$C_2 = 3.71 \times 10^6$ [μm/hr]
	$E_1 = 1.23$ [eV]	
	$E_2 = 1.96$ [eV]	
Steam (@ 640 torr.)	$C_1 = 3.86 \times 10^2$ [μm^2/hr]	
	$C_2 = 1.63 \times 10^8$ [μm/hr]	$C_2 = 0.97 \times 10^8$ [μm/hr]
	$E_1 = 0.78$ [eV]	
	$E_2 = 2.05$ [eV]	

* From S. K. Ghandhi, *VLSI Fabrication Principles,* New York: Wiley, © 1983. Reprinted with permission from John Wiley & Sons, Inc.

and wet oxide growth rates. At a given temperature T, steam oxidation is much faster than that obtained by dry O_2. This faster growth rate is offset by the fact that the presence of excess H_2O found in the steam-grown oxide leads to inferior SiO_2 properties. Critical SiO_2 layers such as those required for MOSFET gate oxides are usually grown in dry oxygen. Examples of oxide thickness as a function of growth time are shown in Fig. 5.2. Increased pressure enhances the oxide growth rate, so *high-pressure* oxidation is commonly used when large x_{ox} values are needed in the chip design. Heavy doping in the silicon also increases the growth rate of an oxide layer.

Thermal oxidation uses silicon atoms at the Si-SiO_2 interface to create molecules of silicon dioxide. The location of the silicon surface will thus be different after the oxide growth is completed. This is illustrated in Fig. 5.3. The thickness x_{Si} of the silicon lost to the oxide growth can be related to the oxide thickness x_{ox} using simple arguments. First, recall that $N_{Si} = 5 \times 10^{22}$ [Si atoms/cm^3] are in the crystal. The silicon dioxide has approximately $N_{SiO_2} = 2.3 \times 10^{22}$ [molecules/cm^3]. Since it takes one atom of silicon to create one molecule of SiO_2,

$$N_{Si} x_{Si} = N_{SiO_2} x_{ox} \tag{5.1-8}$$

expresses the proper balance. Thus,

$$x_{Si} = \frac{N_{SiO_2}}{N_{Si}} x_{ox} \simeq 0.46 x_{ox} \tag{5.1-9}$$

may be used to calculate the amount of silicon consumed in the oxidation process. This will become particularly important when the LOCOS isolation technique is discussed.

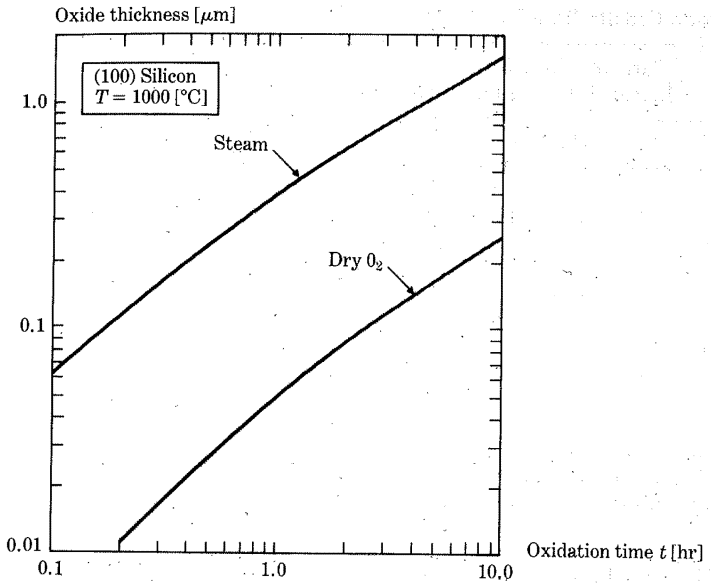


FIGURE 5.2 Example of thermal oxide growth.

It has been found that stray ionic charges, most notably alkali ions such as Na^+, are trapped in thermally grown oxide layers. This leads to problems in MOS threshold voltage variations through the flatband voltage V_{FB}. The amount of ionic charge can be reduced by performing the oxidation in a chlorinated atmosphere, such as that obtained by adding HCl or chlorine gas. The addition of chlorine to the recipe allows for trapping of Na^+ ions into neutral subsystems. This reduces charge at the Si-SiO_2 interface to levels where acceptable MOSFET performance can be attained. Phosphosilicate glass (PSG) may also be used as a method to attract mobile ions to a region of the oxide where they can be neutralized or removed.

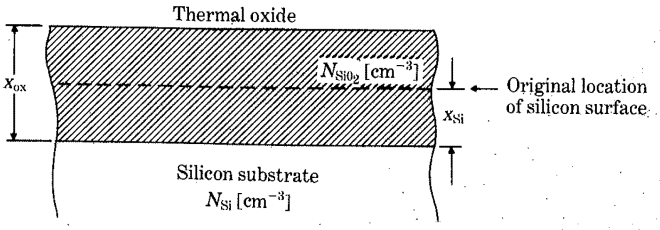


FIGURE 5.3 Geometry for calculation of silicon consumed in the thermal oxidation process.

EXAMPLE 5.1-1

An oxide layer is grown on (111) silicon using steam. The oxidation is performed at 1000 [°C] and lasts for 1 hour. Calculate x_{ox}.

Solution

First find the growth rate constants. For (111) silicon in steam,

$$B = 3.86 \times 10^2 e^{-0.78/(8.62\times10^{-5})(1273)} \simeq 0.316\,[\mu\text{m}^2/\text{hr}],$$

while

$$(B/A) = 1.63 \times 10^8 e^{-2.05/(8.62\times10^{-5})(1273)} \simeq 1.244\,[\mu\text{m/hr}].$$

Dividing,

$$A = \frac{B}{(B/A)} \simeq 0.252\,[\mu\text{m}].$$

Using eqn. (5.1-4) then gives

$$x_{ox} \simeq \frac{(0.252)}{2}\left[\sqrt{1 + \frac{4(0.316)(1)}{(0.252)^2}} - 1\right],$$

so $x_{ox} \simeq 4500\,[\text{Å}]$. ■

5.1.2 Impurity Diffusion

The diffusion of impurity atoms into silicon crystal has been used since the beginnings of the semiconductor industry. This doping technique is based on the fact that dopant atoms gain thermal kinetic energy at elevated temperatures. When this motion is combined with a gradient in the doping density $N(x)$, a net flux of particles F [cm^2-sec]$^{-1}$ is created such that the particles tend to move away from regions of high concentration. Mathematically this motion is described by *Fick's law* in the (one-dimensional) form

$$F = -D\frac{\partial N(x)}{\partial x}, \tag{5.1-10}$$

where D [cm^2/sec] is the *diffusion coefficient*. D depends on the temperature, the dopant species, and the concentration of dopants in the diffusion path. Diffusions are typically performed in the temperature range 900–1100 [°C].

The general properties of diffusive motion are obtained by writing a balance equation for particle densities. Consider the situation shown in Fig. 5.4 where a concentration gradient causes particles to diffuse from region 1 to region 2. Since the number of particles lost from region 1 is (dN/dt), the flow

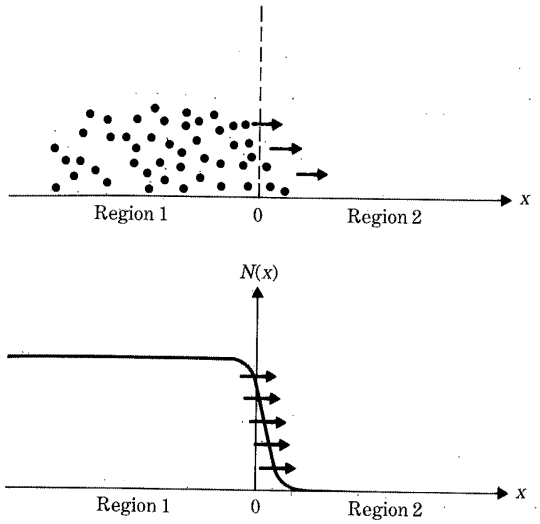


FIGURE 5.4 Particle flow established by a concentration gradient.

is described by

$$\frac{\partial N(x, t)}{\partial t} = -\frac{\partial F(x)}{\partial x}. \tag{5.1-11}$$

Classical diffusion theory assumes that D is not a function of position x. Thus, substituting the flux from eqn. (5.1-10) gives the *diffusion equation*

$$\frac{\partial N(x, t)}{\partial t} = D\frac{\partial^2 N(x, t)}{\partial x^2}. \tag{5.1-12}$$

This partial differential equation can be solved subject to appropriate boundary conditions to determine the impurity profile diffused into the silicon. The diffusion coefficient D in this equation is interpreted as an average value and varies with dopant species and diffusion temperature T [°K]. It may be computed using the formula

$$D = D_0 e^{-E_{av}/kT}, \tag{5.1-13}$$

in which E_{av} is an average activation energy corresponding to the dopant "jumping" from site to site. Approximate values of D_0 and E_{av} are provided in Table 5.2 for B, P, and As.

Dopant profiles are controlled by the manner in which the dopants diffuse into the silicon. Two types of profiles are of particular interest. The first is obtained using a *constant surface concentration* (or *predeposit*) process in which the surface dopant density $N_S = N(0)$ is maintained at a value N_0 [cm^{-3}]. N_0 is

TABLE 5.2 Average Diffusion Coefficients for Common Silicon Dopants

$D = D_0 e^{-E_{av}/kT}$ [cm²/sec]		
Dopant	D_0 [cm²/sec]	E_{av} [eV]
Boron	5.853	3.62
Phosphorus	5.853	3.62
Arsenic	0.091	3.37

known as the *solid solubility limit* and represents the maximum dopant density that can be "dissolved" in the silicon. N_0 depends on the dopant species and temperature as illustrated by the graph of Fig. 5.5. This type of diffusion problem is shown in Fig. 5.6(a). Assuming that the diffusion cycle lasts a time t_p and is carried out at a temperature T_p, the final impurity profile is found to be

$$N(x) = N_0 \operatorname{erfc}\left(\frac{x}{2\sqrt{D_p t_p}}\right), \tag{5.1-14}$$

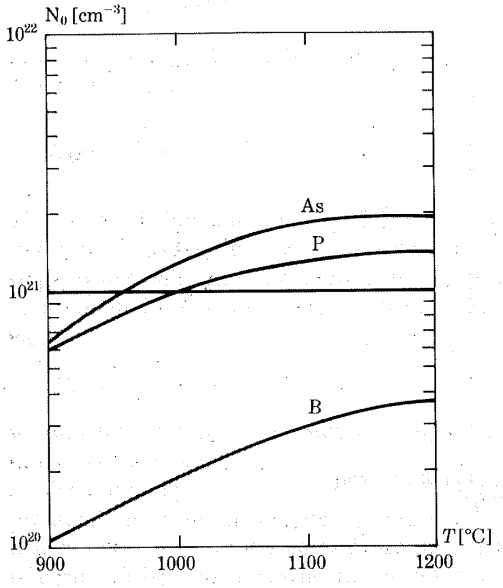
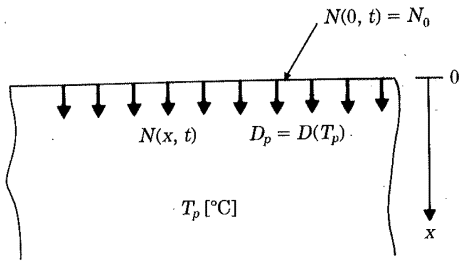
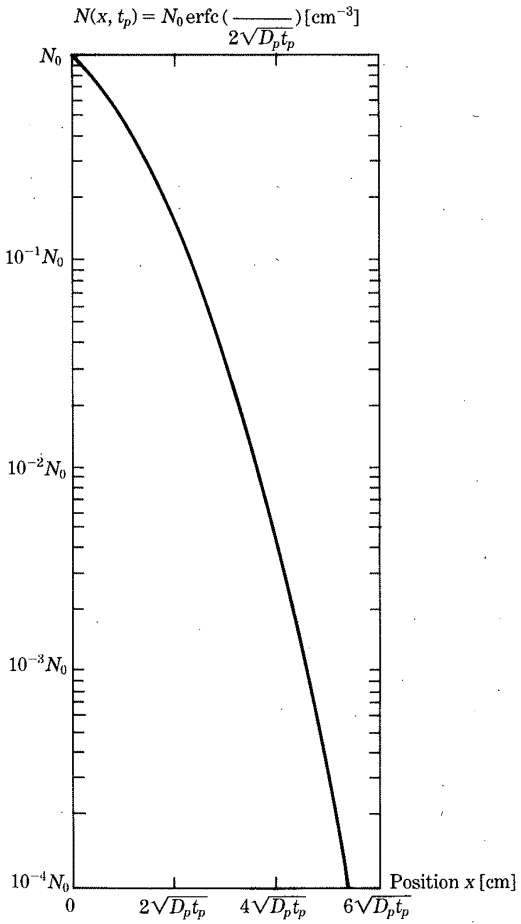


FIGURE 5.5 Solid solubility limits for boron, arsenic, and phosphorus.



Basic diffusion problem
(a)



erfc profile
(b)

FIGURE 5.6 Constant surface concentration diffusion.

where $D_p = D(T_p)$. The *complementary error function* erfc (y) is given by

$$\text{erfc}\,(y) = 1 - \text{erf}\,(y), \tag{5.1-15}$$

where erf (y) is the *error function* defined by the integral representation

$$\text{erf}\,(y) = \frac{2}{\sqrt{\pi}} \int_0^y e^{-u^2}\, du. \tag{5.1-16}$$

A plot of the erfc profile is shown in Fig. 5.6(b). The total number of dopant atoms Q introduced into the wafer during this step is

$$Q = A \int_0^\infty N(x)\, dx = 2N_0 A \sqrt{\frac{D_p t_p}{\pi}}, \tag{5.1-17}$$

with A being the surface area. This type of diffusion is used for shallow doping profiles and also as a first step in deep diffusions.

The second type of diffusion problem is one where the dopants have been introduced into the wafer by a predeposit step and are then subjected to a higher-temperature *drive-in* diffusion at temperature T_d for time t_d. This is shown in Fig. 5.7(a) and is also known as a *limited source* diffusion. Assuming that Q atoms have been left by the predeposit, solving eqn. (5.1-12) gives the Gaussian profile

$$N(x) = \frac{Q}{A\sqrt{\pi D_d t_d}}\, e^{-x^2/4D_d t_d}, \tag{5.1-18}$$

with $D_d = D(T_d)$. Using Q from eqn. (5.1-17) finally yields the result of this *2-step* process as

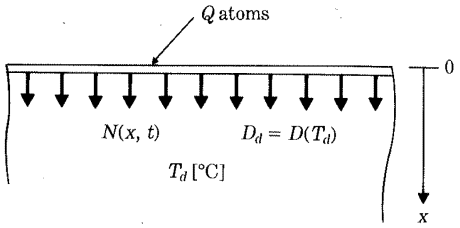
$$N(x) = \frac{2N_0}{\pi} \sqrt{\frac{D_p t_p}{D_d t_d}}\, e^{-x^2/4D_d t_d}. \tag{5.1-19}$$

This profile is illustrated in Fig. 5.7(b) and is valid as long as $(D_d t_d) \gg (D_p t_p)$ is satisfied. Usually the wafer will be subjected to additional heat treatment steps, which will cause the dopants to diffuse deeper into the silicon. This may be accounted for by replacing the factor $(D_d t_d)$ in eqn. (5.1-19) by an *effective* value

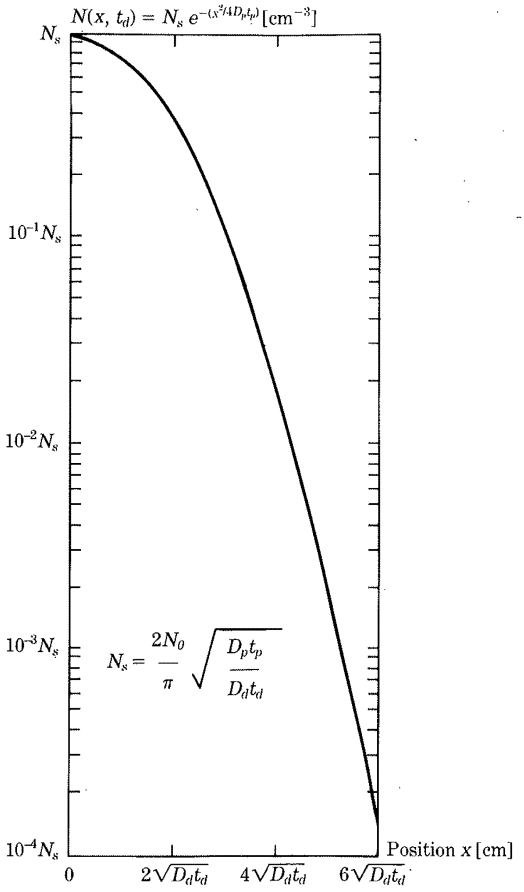
$$\overline{Dt} = D_d t_d + D_1 t_1 + D_2 t_2 + \cdots, \tag{5.1-20}$$

where each term $D_j t_j$ in the sum represents the effect of a separate heating step in the wafer processing.

The doping profiles in eqns. (5.1-14) and (5.1-19) are reasonable approximations as long as the doping levels are low enough to ignore the dependence of D on the carrier densities. High-concentration diffusions used in certain VLSI processes do not result in erfc or Gaussian profiles, and the analysis must be modified to describe these situations. The most general



Basic diffusion process
(a)



Gaussian profile
(b)

FIGURE 5.7 Limited source diffusion.

TABLE 5.3 Concentration-Dependent Diffusion Coefficient Quantities*

$$D = h\left[D_i^o + D_i^-\left(\frac{n}{n_i}\right) + D_i^=\left(\frac{n}{n_i}\right)^2 + D_i^+\left(\frac{p}{n_i}\right)\right]$$

$$D_i^x = D_0^x e^{-E_a^x/kT}$$

Dopant	Boron	Phosphorus	Arsenic
D_0^o	0.091	3.85	0.38
E_a^o	3.36	3.66	3.58
D_0^-	–	4.44	22.9
E_a^-	–	4.0	4.1
$D_0^=$	–	44.2	–
$E_a^=$	–	4.37	–
D_0^+	166.3	–	–
E_a^+	4.08	–	–

D_0^x in [cm²/sec], E_a^x in [eV]

*From R. A. Colclaser, *Microelectronics Processing and Device Design*, New York: Wiley, © 1980. Reprinted with permission from John Wiley & Sons, Inc.

situation is described by writing the diffusion coefficient in the form

$$D = h\left[D_i^o + D_i^-\left(\frac{n}{n_i}\right) + D_i^=\left(\frac{n}{n_i}\right)^2 + D_i^+\left(\frac{p}{n_i}\right)\right], \tag{5.1-21}$$

where h is called the *field enhancement factor*. It accounts for *field-aided* diffusion under the influence of a built-in electric field that is created by the separation of electrons from their host dopant atoms. h has a value between 1 and 2. n and p are the carrier densities, while n_i is the intrinsic concentration at the diffusion temperature. The quantity D_i is the intrinsic diffusion coefficient for the *x charge-state vacancy*. Intrinsic values are computed from

$$D_i^x = D_0^x e^{-E_a^x/kT}, \tag{5.1-22}$$

where E_a is the activation energy for the state. Table 5.3 provides values of D_0 and E_a for B, P, and As. Note that only certain terms in eqn. (5.1-21) contribute to D for a given dopant.

Boron is a p-type (acceptor) dopant that exhibits neutral and positive charge-state vacancy interactions. The diffusion coefficient has the form

$$D \simeq D_i^o + D_i^+\left(\frac{p}{n_i}\right), \tag{5.1-23}$$

where the field enhancement factor has been set to unity. It is convenient to define an average intrinsic diffusion coefficient by

$$D_i = D_i^o + D_i^+. \tag{5.1-24}$$

A constant surface concentration boron diffusion is then found to be described by the profile

$$N(x, t_p) \simeq N_0(1 - Y^{2/3}), \tag{5.1-25}$$

where

$$Y = \left(\frac{n_i x^2}{6N_0 D_i t_p}\right)^{3/2}. \tag{5.1-26}$$

This type of diffusion result is shown in Fig. 5.8 where it is compared to the

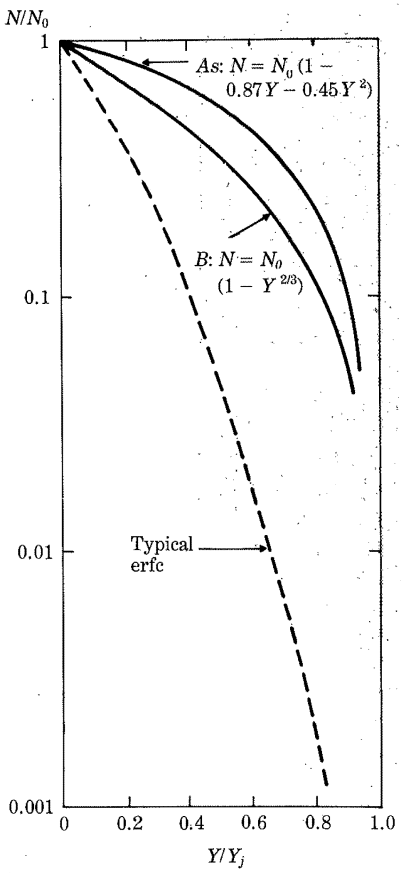


FIGURE 5.8 High-density constant surface concentration diffusion profile comparison.

erfc distribution. The diffusion edge is much sharper than that predicted by the classical theory. When this type of boron diffusion is performed into an n-type background, a pn junction is formed. The sharpness of the curve allows the location x_j of the junction to be found by setting $N(x_j, t_p) = 0$, i.e., $Y = 1$. Thus,

$$x_j \simeq 2.45\sqrt{D_i t_p}\sqrt{\frac{N_0}{n_i}} \tag{5.1-27}$$

allows for an estimate of the diffusion depth.

High-density arsenic (n-type) diffusions are characterized by

$$D = h\left[D_i^o + D_i^-\left(\frac{n}{n_{ie}}\right)\right], \tag{5.1-28}$$

showing interactions with neutral and single-negative charge-state vacancies. The *effective intrinsic carrier density* n_{ie} is introduced to account for *bandgap narrowing* from the high doping levels. The effective density is computed from

$$n_{ie}^2 = n_i^2 e^{\Delta E_g/kT} > n_i^2 \tag{5.1-29}$$

with ΔE_g the reduction in the energy gap. Figure 5.9 shows n_i and n_{ie} over the standard range of diffusion temperatures.

It is found that the field enhancement factor for As is approximately

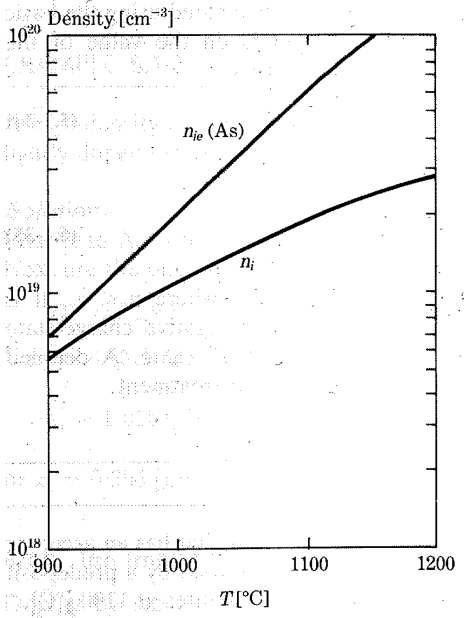


FIGURE 5.9 n_i and n_{ie} as functions of temperature.

$h = 2$. Using this with the assumption that $(n/n_{ie}) \gg 1$ reduces the diffusion coefficient to

$$D \simeq 2D_i^-\left(\frac{n}{n_{ie}}\right). \tag{5.1-30}$$

A constant surface concentration diffusion lasting a time t_p results in a doping profile that is approximated by

$$N(x, t_p) = N_0(1 - 0.87Y - 0.45Y^2), \tag{5.1-31}$$

with Y the dimensionless variable

$$Y = \sqrt{\frac{n_{ie}x^2}{8N_0D_i^-t_p}}. \tag{5.1-32}$$

This type of distribution is also plotted in Fig. 5.8. When As is diffused into a lightly doped p-type sample, the pn junction depth may be calculated by setting $N(x_j, t_p) = 0$. Solving the quadratic results in

$$x_j \simeq 1.616\sqrt{2D_i^-t_p\left(\frac{N_0}{n_{ie}}\right)}, \tag{5.1-33}$$

which provides a good estimate of the diffusion depth.

The diffusive behavior of high-density phosphorus diffusions is quite complicated and is termed *anomalous* as it cannot be explained using the basic theory. The form of the diffusion coefficient depends on the value of the electron concentration with respect to a reference density

$$n_e \simeq 4.65 \times 10^{21}e^{-0.39/kT}. \tag{5.1-34}$$

In regions near the surface where $n > n_e$,

$$D = h\left[D_i^o + D_i^=\left(\frac{n}{n_{ie}}\right)^2\right] \tag{5.1-35}$$

is valid. However, in the deeper regions of the crystal where $n < n_e$, it is found that phosphorus has neutral and complex single-negative charge-state interactions. This changes D to what is termed the "tail" value. A detailed discussion of this topic is beyond the scope of the present treatment.

EXAMPLE 5.1-2

A 2-step phosphorus diffusion is made into a p-type wafer that has an acceptor doping density of $N_a = 10^{15}$ [cm^{-3}]. The diffusion is described by a predeposit for 30 minutes at 900 [°C] followed by a drive-in for 60 minutes at 1100 [°C].

Calculate the depth x_j of the pn junction formed by this process.

Solution
First read $N_0 = 6 \times 10^{20}$ [cm^{-3}] for the solid solubility. Next compute the diffusion coefficients at the two temperatures:

$$D_p = 5.853e^{-3.62/(8.62\times10^{-5})(1173)} \simeq 1.66 \times 10^{-15} \text{ [cm}^2\text{/sec]},$$

$$D_d = 5.853e^{-3.62/(8.62\times10^{-5})(1373)} \simeq 3.15 \times 10^{-13} \text{ [cm}^2\text{/sec]}.$$

Then use eqn. (5.1-19) to find the doping profile:

$$N(x) = \frac{2(6 \times 10^{20})}{\pi}\sqrt{\frac{(1.66 \times 10^{-15})(30)}{(3.15 \times 10^{-13})(60)}}\, e^{-x^2/4(3.15\times10^{-13})(60)(60)}.$$

This gives

$$N(x) \simeq 1.97 \times 10^{19} e^{-x^2/0.454} \text{ [cm}^{-3}\text{]},$$

where x is in units of [μm].

To calculate the junction depth, just look for the point where the donor doping exactly equals the p-type acceptor level N_a, i.e.,

$$N(x_j) = N_a.$$

Solving,

$$x_j \simeq \sqrt{0.454 \ln\left(\frac{1.97 \times 10^{19}}{10^{15}}\right)} \simeq 2.12 \text{ [}\mu\text{m]}.$$ ■

EXAMPLE 5.1-3

A high-density arsenic diffusion is performed at 1000 [°C] for 30 minutes into a lightly doped p-type wafer. Calculate the junction depth x_j.

Solution
First read $N_0 \simeq 6.5 \times 10^{20}$ [cm^{-3}] and $n_{ie} \simeq 2 \times 10^{19}$ [cm^{-3}] from the charts. Next use the data in Table 5.3 to calculate

$$D_i^- = 22.9e^{-4.1/(8.62\times10^{-5})(1273)} \simeq 1.36 \times 10^{-15} \text{ [cm}^2\text{/sec]}.$$

Equation (5.1-33) then gives

$$x_j \simeq 1.616\sqrt{2(1.36 \times 10^{-15})(30)(60)\frac{6.5 \times 10^{20}}{2 \times 10^{19}}},$$

or $x_j \simeq 0.204$ [μm]. ■

5.1.3 Ion Implantation

Doping by ion implantation has already been mentioned in Section 1.1 for setting MOSFET threshold voltages. An ion implanter is an apparatus that first

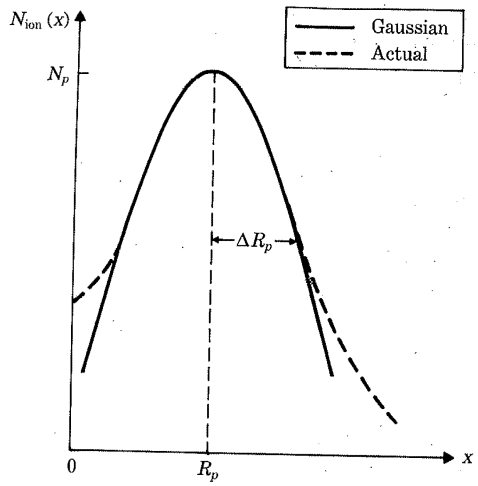


FIGURE 5.10 Gaussian approximation for ion implant profile.

ionizes a dopant source and then accelerates the ions to high energies. The ions form a beam that is focused to scan over the wafer. The dopant ions are literally smashed into the substrate. This results in a doping profile that is approximated to first order by the Gaussian

$$N_{\text{ion}}(x) = \frac{D_I}{\sqrt{2\pi}(\Delta R_p)} e^{-(1/2)[(x-R_p)/\Delta R_p]^2}. \tag{5.1-36}$$

The projected range R_p represents the average penetration depth of the implanted ions. R_p increases with increasing energies and is also a function of the dopant mass. The straggle (ΔR_p) is the standard deviation associated with the implanted depth and depends on both the incident ion energy and the ion mass relative to the silicon atomic mass. The general shape of the Gaussian approximation is illustrated in Fig. 5.10, where it is compared to a more realistic implant profile.

The number of ions implanted into a wafer is specified by using the dose D_I [cm^{-2}] defined as

$$D_I = \int_{-\infty}^{+\infty} N_{\text{ion}}(x)\, dx = N_p(\Delta R_p)\sqrt{2\pi}, \tag{5.1-37}$$

where N_p represents the peak concentration (at $x = R_p$) in units of [cm^{-3}]. The dose is a carefully controlled quantity and can be measured using the *current integrator* technique shown in Fig. 5.11. This approach to *dosimetry* measures the current supplied to the wafer to offset the charge imbalance created by the ion beam. To understand the system, let I represent the beam

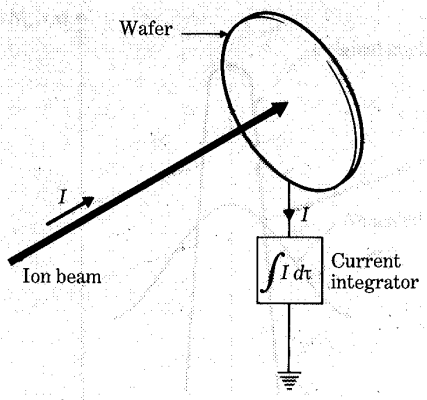


FIGURE 5.11 Current integrator dosimetry arrangement.

current so that

$$\Delta Q_o = \int_0^{t_I} I(\tau)\,d\tau \tag{5.1-38}$$

is the total charge delivered to the wafer during an implant time t_I. The charge on an individual ion can be written as $(\pm sq)$, where $s = 1$ or 2 depending on whether the ions are singly or doubly charged. The total number of ions implanted into the wafer is thus given by $(\Delta Q_o/sq)$. Denoting the implanted area by A yields the dose expression

$$D_I = \frac{1}{sqA}\int_0^{t_I} I(\tau)\,d\tau. \tag{5.1-39}$$

This shows that the implant dose can be computed directly from knowledge of the beam current. Dose measurements typically exhibit less than 1% error, which serves to make ion implantation one of the most carefully controlled processes available in silicon wafer fabrication technology.

Ion implant profiles result from the statistical nature of the *ion-stopping process*. An incident ion is assumed to have a mass M_I and energy E_I, as shown in Fig. 5.12. The ion loses energy to the crystal by collisions with electrons and silicon nuclei. The maximum energy E_T transferred to a *target* mass M_T (and hence lost by the ion) is found to be

$$E_T = 4\frac{M_I M_T}{(M_I + M_T)^2}E_I, \tag{5.1-40}$$

where M_T is either the electron mass m or the nuclear mass M_{Si}, depending on the event. Coulombic forces also contribute to the ion energy loss. The overall effects are contained in the *stopping power* S [eV/cm] such that

$$S = S_n(E_I) + S_e(E_I) \tag{5.1-41}$$

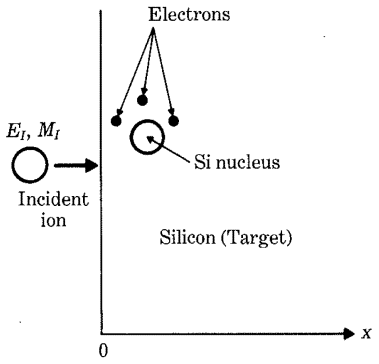


FIGURE 5.12 Basic ion-stopping problem.

gives the nuclear and electronic contributions. The ion energy loss is then described by

$$-\frac{dE_I}{dx} = N_{Si}[S_n(E_I) + S_e(E_I)]. \tag{5.1-42}$$

This may be used to find the implanted distance R for an ion by integrating in the form

$$R = \int_0^R dx = -\int_{E_I}^0 \frac{dE_I}{N_{Si}[S_n(E_I) + S_e(E_I)]}. \tag{5.1-43}$$

The limits on the energy integral correspond to having an incident ion energy E_I at $x = 0$ and to an ion at rest ($E = 0$) at $x = R$. The projected range R_p is the average value of R for all of the ions. The straggle ΔR_p represents the statistical variations in the ion penetration distances.

The preceding brief description of the ion-stopping process illustrates two important problems introduced by the implant. First, the collisions with the targets will tend to knock Si atoms out of their equilibrium lattice sites, which causes cracks and defects in the crystal. Second, the ions will be randomly distributed in the lattice and cannot act as substitutional impurities until the implant is *activated*. This requires that the dopants find positions corresponding to normal lattice sites in the crystal. To solve these two problems, the wafer is subjected to an *annealing* step that simultaneously repairs the damage and activates the implant.

Thermal annealing consists of heating the wafer for a short time to allow for diffusion of both the dopants and the silicon atoms. Assuming that this is performed at a temperature T_A for a time t_A, the thermal anneal has the effect of changing the straggle to

$$(\Delta R_p) \rightarrow \sqrt{(\Delta R_p)^2 + 2D_A t_A}, \tag{5.1-44}$$

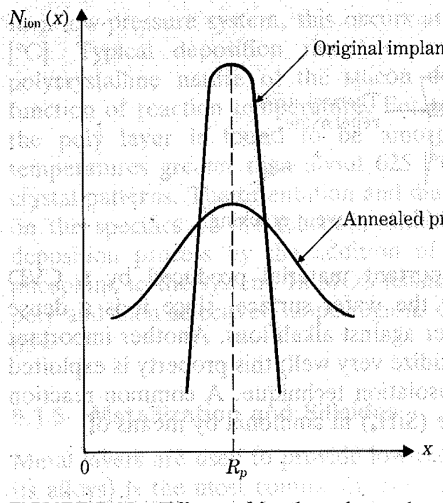


FIGURE 5.13 Effects of implant thermal annealing.

where $D_A = D(T_A)$ for the implanted species. A typical anneal cycle might be 30 minutes at 1000 [°C]. Using this in eqn. (5.1-36) gives the final implant form

$$N_{ion}(x) = \frac{D_I}{\sqrt{2\pi}\sqrt{(\Delta R_p)^2 + 2D_A t_A}} \exp\left[-\frac{1}{2}\left(\frac{x - R_p}{\sqrt{(\Delta R_p)^2 + 2D_A t_A}}\right)^2\right]. \quad \textbf{(5.1-45)}$$

As shown by the plots in Fig. 5.13, the thermal anneal causes the dopants to diffuse away from the projected range point R_p. This spreads the distribution and reduces the peak concentration. A more accurate analysis would account for reflections and out-diffusions at the silicon surface.

Rapid anneal techniques have recently become of interest. This approach involves heating the wafer to high temperatures for very short time periods. An example of a rapid anneal technique is *laser annealing*, which employs a pulsed high-energy laser beam focused to scan the wafer. This induces localized melting and recrystallization. The primary advantage of this approach is that the annealed profile is very close to the original implant distribution, i.e., the spread described by eqn. (5.1-44) is not found for this type of annealing. The main disadvantage is that the *wafer throughput* (the number of wafers processed per hour) is greatly reduced.

5.1.4 Chemical Vapor Deposition

Chemical vapor deposition (CVD) is used to obtain layers of various materials in silicon wafer fabrication. As implied by its name, CVD produces layers by using chemical reactions that deposit the desired material onto the substrate. The reaction kinetics involved in this type of process depend on the temperature, the pressure, and the reactant species.

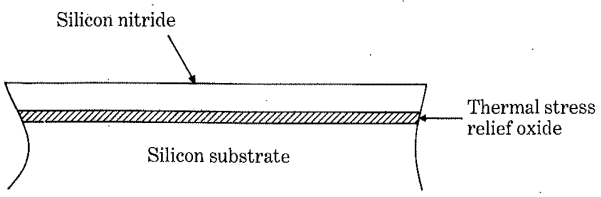


FIGURE 5.14 Nitride with stress-relief oxide to prevent cracking.

Silicon nitride (Si_3N_4) is an important material produced by a CVD process. Nitride is used to *passivate* the wafer surface, since it is a dense dielectric that forms an effective barrier against alkali ions. Another important property of Si_3N_4 is that it does not oxidize very well; this property is exploited to provide the basis for the LOCOS isolation technique. A common reaction for producing nitride layers uses silane (SiH_4) in ammonia by means of

$$3SiH_4 + 4NH_3 \rightarrow Si_3N_4 + 12H_2.$$

The reaction temperature for this process is on the order of 700 [°C]. One problem that arises is that nitride layers exhibit very high tensile stress, which can lead to cracks when the nitride is deposited directly onto a silicon surface. To avoid this situation, a thin oxide layer is usually grown on a silicon surface prior to the nitride deposition, as shown in Fig. 5.14. This *stress relief oxide* tends to relieve some of the surface tension and allows for reliable coverage by the nitride.

CVD oxide layers are required whenever a silicon surface is not available for a thermal oxide growth. One of the more common reactions used for this purpose is the pyrolysis of silane in oxygen as described by

$$SiH_4 + 2O_2 \rightarrow SiO_2 + 2H_2O \qquad (600\text{–}1000\ [°C]),$$

or

$$SiH_4 + O_2 \rightarrow SiO_2 + 2H_2 \qquad (300\text{–}500\ [°C]).$$

The lower-temperature reaction produces hydrogen as a by-product and results in higher-quality films than those obtained using the high-temperature water-producing process. Phosphosilicate glass (PSG) can be deposited by simultaneous pyrolysis of silane and phosphine (PH_3). Phosphine reacts with oxygen according to

$$2PH_3 + 4O_2 \rightarrow P_2O_5 + 3H_2O.$$

The P_2O_5 is then incorporated into the silica during the deposition process. PSG has a lower tensile stress than pure silica, which tends to make PSG more reliable when many thermal cycles are required in the process flow.

Polycrystalline silicon (poly) may be deposited using the direct pyrolysis of silane

$$SiH_4 \rightarrow Si + 2H_2.$$

In a low-pressure system, this occurs at a reaction temperature of about 600 [°C]. Typical deposition rates are on the order of 200 [Å/min]. The polycrystalline nature of the silicon deposited in this manner is a strong function of reaction temperature. For temperatures less than about 575 [°C], the poly layer is found to be amorphous. In contrast, layers grown at temperatures greater than about 625 [°C] tend to exhibit localized columnar crystal patterns. The orientation and diameter of these columns are dependent on the specifics of the reaction conditions. Poly may be doped during the deposition process by the addition of diborane (B_2H_6), arsine (AsH_3), or phosphine to the system. In MOS technology it is common to deposit intrinsic poly and then selectively dope regions of the layer during the normal process flow.

5.1.5 Metallization and Silicides

Metal layers are used to provide low-resistance interconnects. Aluminum (and its alloys) is the most commonly used metal for this application, with a room temperature resistivity of about 2.7 [mΩ-cm]. Gold is also useful for certain applications. Metallic layers of this type are obtained by first melting and then evaporating the metal in a closed vacuum system. The solid samples are heated using standard resistance wiring, electron beams, or rf-induction heating techniques. Placing the wafers into the evaporant flux allows for a complete and (almost) uniform coverage of the surface. Condensation of the metal onto the wafer is attained by keeping the temperature of the wafer lower than that of the evaporant.

Refractory (high-temperature) metals are finding increased use in VLSI structures. The main advantage of using a refractory metal is that low-resistance patterned interconnect levels can be formed at any time in the chip process flow. The high melting point associated with these metals keeps the pattern intact even if the wafer is subjected to additional heat treatments. (Al has a low melting point—660 [°C]—and can be used only after all the high-temperature cycles are completed.) This property allows for multiple-level interconnects to be created, which aids in the layout of complex chips. Refractory metals of current interest include titanium (Ti), platinum (Pt), molybdenum (Mo), and tungsten (W). These may be evaporated using electron-beam heating or may be *sputtered* onto the wafer. In sputtering, an inert gas such as Ar is ionized and used to bombard an electrode that is coated with the desired substance. The collisions knock the atoms off the electrode, creating a particle flux directed toward the wafer.

Silicides are compounds formed using silicon and a refractory metal. These have been introduced for the same reason as the elemental refractory metals, i.e., to create high-conductivity paths that can withstand high-temperature processes without melting. Some of the better-studied silicides are $TiSi_2$, $PtSi_2$, TaSi, and WSi_2. In this group, titanium disilicide ($TiSi_2$) has shown the lowest resistivity, with about $\rho = 2.5$ [mΩ-cm] for cosputtered samples.

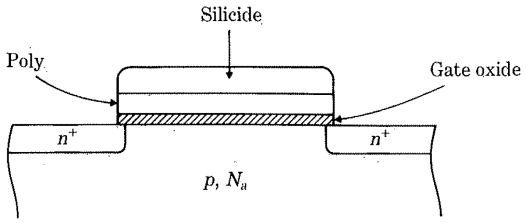


FIGURE 5.15 A polycide gate MOSFET.

Polycides are simply combinations of polysilicon and silicides. A typical polycide consists of a lower level of poly with an upper coating of silicide. These are particularly useful because they exhibit the best features of both polysilicon and silicides, including good stability, adherence, and coverage (from the poly) and a high conductivity path (from the silicide). A MOSFET fabricated with a polycide gate is shown in Fig. 5.15. It should be noted that the threshold voltage of this device is dependent on the specifics of the layering used for the gate. If a thick poly layer (>1500 [Å]) is used, then the gate threshold voltage will be the same as if only poly were used. However, if the poly layer is thin, a shift in the gate-substrate work function Φ_{GS} will occur because of the silicide. This must be accounted for in the threshold voltage adjustment implant.

5.2 Lithography

An integrated circuit may be viewed as a set of patterned layers. Each layer is a material with specific electrical properties. For example, n^+ regions, silicon dioxide, and aluminum form layers of doped semiconductor, insulator, and metal, respectively. In general, a layer must be patterned before the next layer of material is applied. A device such as a MOSFET is a subset of the patterned chip layers. Device characteristics result from the 3-dimensional structure created by both the layering order and the geometric pattern on each layer.

The process used to transfer a pattern to a layer of an integrated circuit is called *lithography*. Since each layer has its own distinct patterning requirements, the lithographic sequence must be repeated for every layer used to build the chip. The main requirements for a lithographic process are (a) a *mask* or *reticle* that is a "master copy" of the desired pattern; (b) an *exposure* or *printing* system that allows the transfer of the pattern to the wafer; and (c) a *resist* material that coats the wafer and is itself patterned by the printing system as the first step in creating the patterned layer.

5.2.1 Basic Pattern Transfer

The easiest way to illustrate the important aspects of a lithographic process is through a specific example. The one chosen here is basic *optical* lithography

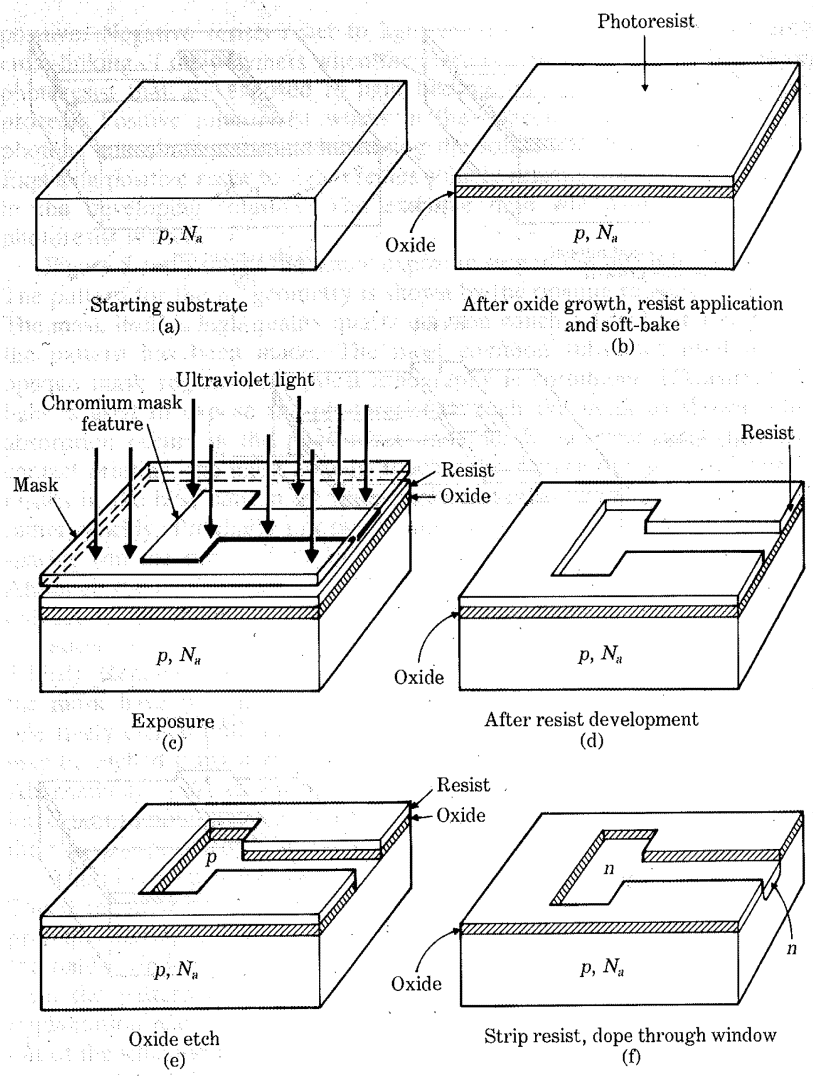


FIGURE 5.16 Example process for optical lithography.

with *contact/proximity* printing. The process will be discussed in conjunction with the wafer fabrication steps in Fig. 5.16. This set of drawings will show how a patterned n^+ region is created in a p-type substrate such that the n^+ *well* is accessed by a metal interconnect layer.

The starting point for the example is the section of p-type wafer with constant doping N_a shown in Fig. 5.16(a). The wafer is first subjected to a

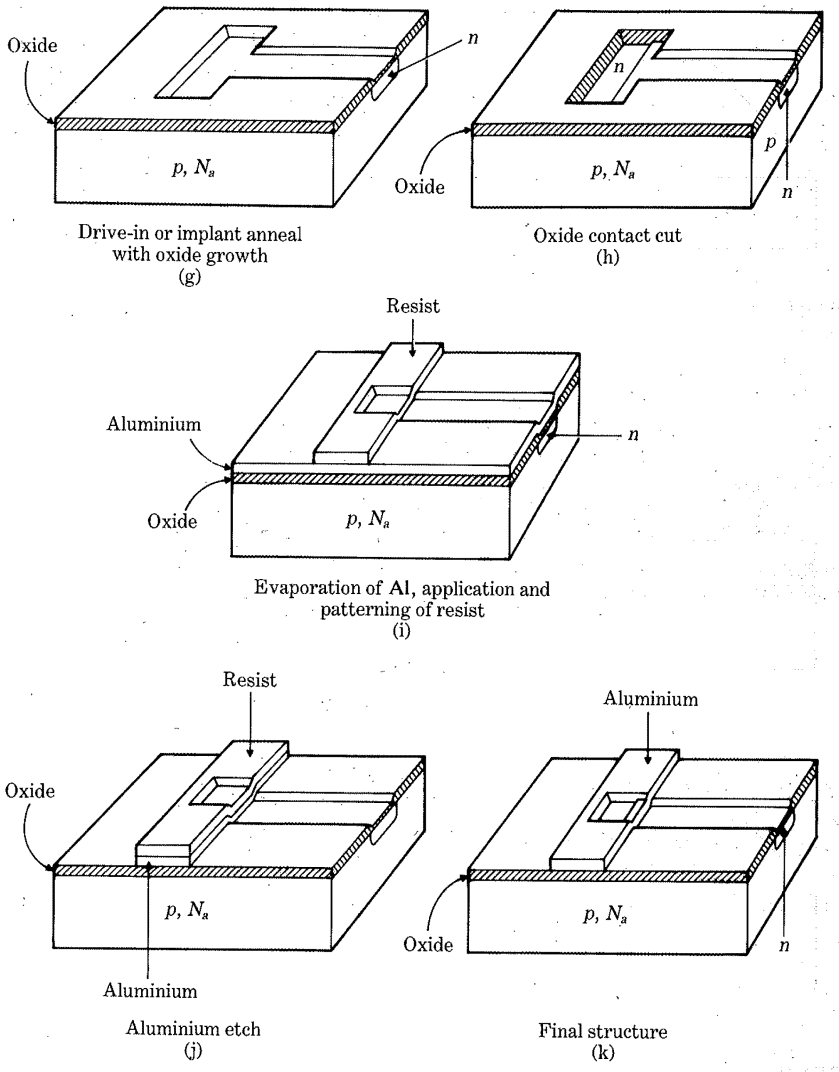


Figure 5.16 (*contd.*)

thermal oxide growth and is then coated with a layer of *photoresist.* At this point, the wafer has the layering of Fig. 5.16(b). Photoresist is a light-sensitive organic polymer (generally a type of plastic) that is initially in a liquid state when it is applied to the wafer. The wafer is spun to allow for an even coating of resist, and it is then subjected to a *soft-bake* heating step that solidifies the resist enough for handling. Photoresist is available in two types, negative and

positive. Negative resists react to light by absorbing photons, which creates cross-linking of the polymers when the resist is developed. Regions of negative photoresist that are exposed to light become hardened after the developing process. Positive photoresist works in the opposite manner, with absorbed photons contributing toward increasing the solubility of the resist base *resin*. Exposing positive resist to light creates soluble regions that are dissolved away in the developing solution. The example here will assume that negative photoresist is used.

Figure 5.16(c) shows the actual exposure step in the first lithographic cycle. The pattern for the n^+ geometry is shown by the opaque regions on the mask. The mask itself is high-quality quartz glass on which a 1-to-1 (or 1 ×) copy of the pattern has been made. The most common substance used to create opaque mask regions for optical lithography is chromium. Ultraviolet (UV) light is used to expose the photoresist through the mask as shown. Photon absorption occurs in the photoresist underneath the clear mask regions. In contact printing, the mask actually touches the surface of the resist layer. This results in the best pattern transfer resolution but tends to wear out the mask rather quickly. Proximity printing solves this problem by maintaining a small spacing (on the order of 20–40 [μm]) between the mask and wafer surface. Although the mask does not degrade as rapidly, the resolution of the image is reduced from that obtained by contact printing due to diffraction.

After exposure, the resist is developed to obtain the patterning as in Fig. 5.16(d). Regions of negative resist that were exposed to the UV light through the mask have become hardened. This allows for the silicon dioxide to be selectively etched into the pattern, illustrated by Fig. 5.16(e). The silica layer may be etched using a wet chemical bath of buffered hydrofluoric (HF) acid. Alternatively, a *dry etch* process based on a gaseous plasma may be used. It is important to note that this step has opened a patterned oxide window down to the bare p-type silicon substrate.

The next step in the processing is to strip the photoresist off the wafer. This is required because resist generally has a low melting temperature, which prohibits leaving it on during high-temperature cycles. The pattern provided by the oxide window is then used to selectively dope the wafer with donors to form the patterned n-type well in the p-type substrate. Either diffusion or implantation may be used since the oxide acts as a mask to keep the dopants out of the silicon regardless of the doping technique employed. The oxide itself becomes doped during this process but is assumed to be thick enough to keep the dopants from reaching the silicon surface. The resulting n-type well that is created in this step is shown in Fig. 5.16(f).

After the initial doping is completed, the wafer is heated for either a drive-in (for a diffusion) or an anneal (with an ion implant). The atmosphere is assumed to be nitrogen and oxygen, so an oxide layer is grown over the entire surface. This is drawn in Fig. 5.16(g). The oxide is introduced at this point as an insulator between the semiconductor and the metal layer.

Access to the n^+ well is obtained by another lithography step, which

results in the oxide contact window shown in Fig. 5.16(h). Aluminum is then evaporated over the entire surface of the wafer, and the desired metal pattern is formed on the resist layer with the next lithography sequence. The wafer at this point in the processing appears as illustrated by Fig. 5.16(i). Etching the aluminum leaves the patterned layers shown in Fig. 5.16(j). The last step in this example is to strip the resist, which leaves the final structure shown in Fig. 5.16(k).

This example demonstrates that the masking process is repeated for every layer that needs patterning. In particular, masks were used to pattern the n^+ doping, the oxide contact cut, and the metal geometry. Owing to the importance of obtaining patterned layers that are accurate images of the mask, good lithography is crucial for making integrated circuits. Many variations in the lithography sequence are possible. Although space prohibits a detailed discussion of the subject, some of the possibilities are discussed.

The first to be examined is an exposure technique known as *projection* printing. In this approach, the mask is kept some distance away from the wafer, and the pattern transfer is effected by optically projecting the image through a lens system. One example of this type of printing system is the *wafer stepper* shown in Fig. 5.17(a). The layer patterning is contained on the reticle, which is typically a 5× or 10× enlargement of a single die area. Pattern transfer is accomplished by optically reducing the image to the correct dimensions while simultaneously focusing it onto a specified region of the wafer. After the portion of the wafer is exposed, the system "steps" the image

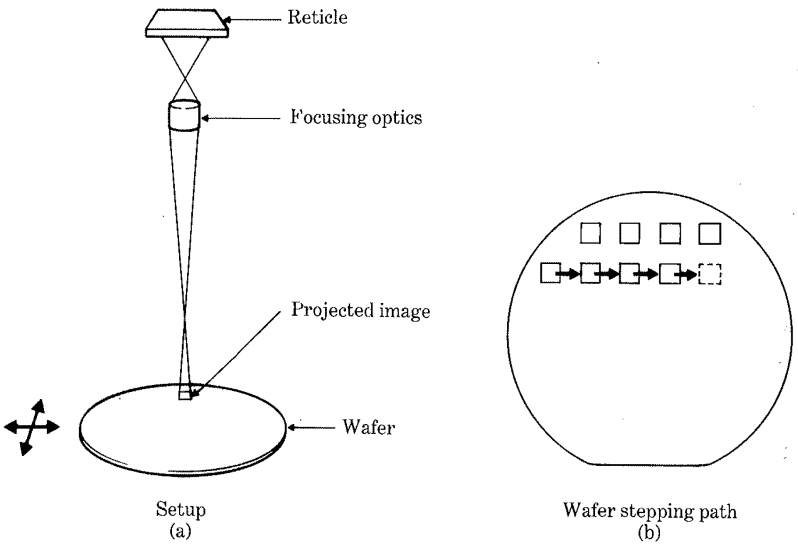


FIGURE 5.17 Wafer stepper projection printing system.

to the next die site, as schematically illustrated by Fig. 5.17(b). This stepping process is continued until the wafer is completely exposed.

The direct wafer stepping technique is useful because it allows for wafer surface variations due to warping. This is particularly important as wafer diameters increase. The use of a 5× or 10× reticle pattern allows for larger tolerances in the pattern definition, making it more accurate than that possible with a 1× system. Of course, diffraction effects compensate for some of this because of the nature of the projection system. Another advantage of this technique is that the pattern transfer is not greatly affected by dust particles because the rectile is shielded with a sheath of clear nitrocellulose (or similar material) called a *pellicle*. The system is arranged so that any dust accumulation is restricted to the pellicle surface. Image control is maintained by ensuring that the distance between the reticle and the pellicle is sufficient to keep the dust particles out of the focusing range of the lens system. Consequently, the dust particles do not image onto the wafer. The main disadvantage of the wafer stepper is that the wafer throughput is low compared with that of a single-exposure process.

Negative photoresists are based on light-sensitive organic polymeric resins to which photosensitizing agents have been added to aid in the photon absorption process. Typical negative-working resins are polyvinyl cinnamate and cyclized polyisoprene. The resin itself must have the basic properties that it can adhere to the desired surface and can be developed into a hardened film that is resistant to the basic layer etchants. The specific sensitizer used to create a working negative resist depends on the resin system since the two must be compatible. Negative photoresist can be made very sensitive to light, allowing for rapid exposure. This makes negative photoresists useful for high wafer throughput. However, they have the undesirable property of swelling during the development process because the developing agent is absorbed by the polymer. Owing to the distortion introduced by the swelling, negative photoresists tend to be limited to minimum linewidths on the order of 2 [μm].

Positive photoresists do not absorb the developing solution and are therefore the prime candidates for fine-line optical VLSI lithography. The polymeric chemistry associated with positive photoresists is quite different from that described above for negative-working resists. Resins used in positive photoresists are not generally sensitive to light but are chosen for their basic characteristics, such as solubility and adherence. The exposure properties of positive resists arise from the reaction of the added light sensitizers with the base resin chains. Positive photoresists tend to require longer exposure times than those needed for negative resists. Consequently, improved resolution capabilities are gained at the expense of reduced wafer throughput.

X-ray lithography is similar to the optical UV technique described above except that X-rays are used as the exposure source. This approach is of interest because the shorter wavelengths of X-rays allow for finer resolution in the pattern transfer process while still maintaining high wafer throughput rates.

The most difficult aspect of X-ray lithography is mask making. The requirements are somewhat different from those used in optical lithography in that the opaque regions of an X-ray mask are made to absorb (rather than reflect) the radiation. Gold is typically used as an absorber for this type of mask. The walls for the gold must be maintained perpendicular to the mask surface to ensure good resolution.

Electron-beam lithography offers the highest-resolution pattern formation by directly writing onto the wafer. The electron beam is columnated to form the desired *spot size,* and the beam is then used to scan the wafer surface, as shown in Fig. 5.18(a). The pattern information is contained in the e-beam scan control system, which selectively turns the beam on and off depending on the wafer coordinates of the beam as it scans the surface. Since beam diameters as small as 0.01 [μm] are possible, direct e-beam writing allows for extremely fine-line patterning. The major drawback of this technique is that the wafer throughput is low because of the time required to scan the wafer at a rate slow enough to expose the resist. The throughput can be increased by enlarging the spot size, but this obviously reduces the resolution. A typical production line system providing a minimum linewidth of 2 [μm] might use an e-beam as large as 0.5 [μm]. Other problems are that the circular beam shape makes it difficult to form right-angle (90°) turns in the pattern and that electron backscattering, illustrated in Fig. 5.18(b), tends to widen the actual exposure track.

5.2.2 Lateral Doping Effects

Doping through an oxide window results in a doped region that is larger than the window itself. This occurs because of *lateral doping* effects in which there is a "spreading" of the dopants outward from the window edge in directions parallel to the silicon surface. Both diffused and implanted patterns exhibit lateral spreading. The MOSFET overlap capacitances C_{ols} and C_{old} discussed in Section 4.6, are direct consequences of this phenomenon.

Lateral diffusion is illustrated in Fig. 5.19(a). It is seen that the oxide window serves as a mask for the patterning but that dopants at the edge of the oxide window diffuse outward. This results in the diffused well geometry shown in Fig. 5.19(b). The amount of lateral spreading is usually characterized by the lateral junction distance y_j. This can be quite large, with $(y_j/x_j) = 0.8$ a typical order of magnitude for a deep 2-step Gaussian profile. The actual value of y_j is dependent on the background doping N_{BC} of the substrate and the specifics of the diffusion cycle.

Lateral doping in ion implantation initially results from ion scattering outside the oxide "window plane" during the ion-stopping process. This is shown in Fig. 5.20(a). Thermal annealing of the implant will induce further spreading because of lateral diffusive motion during the heat treatment. The final width of the implant is thus larger than the oxide window, as depicted in Fig. 5.20(b). Lateral doping from an ion implant is generally much less than

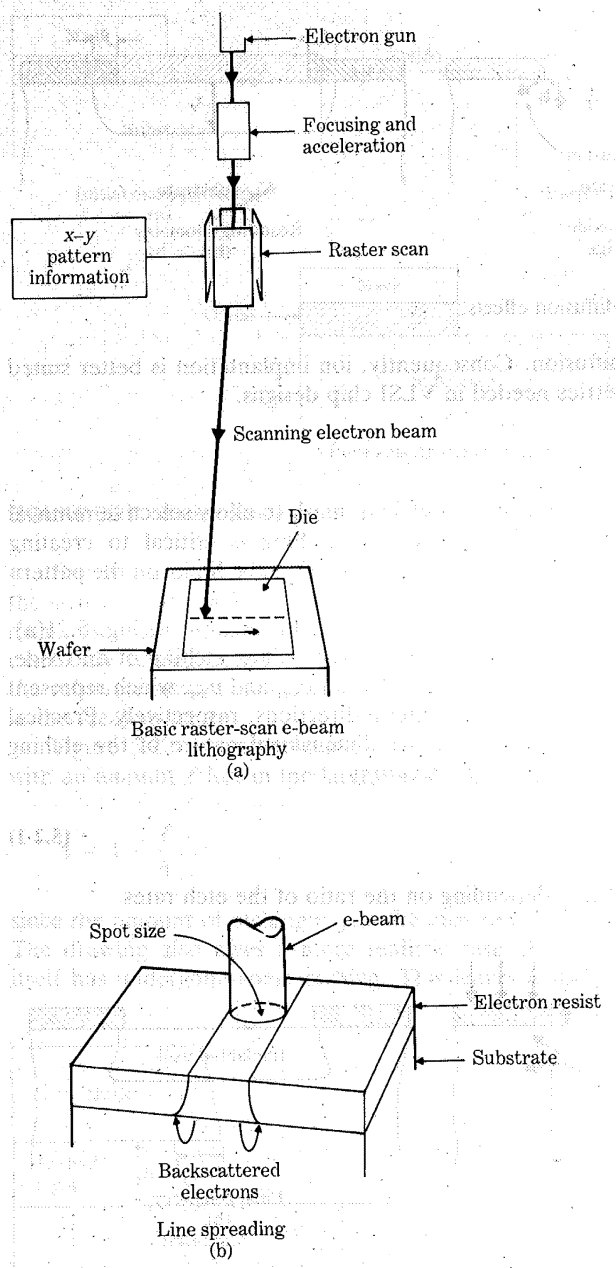


FIGURE 5.18 Electron-beam lithography.

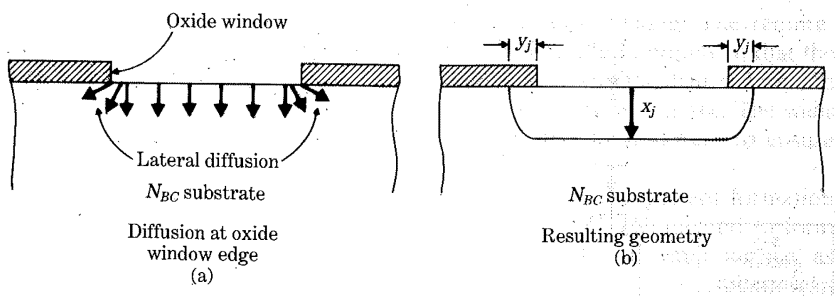


FIGURE 5.19 Lateral diffusion effects.

that for an impurity diffusion. Consequently, ion implantation is better suited for the fine-line geometries needed in VLSI chip designs.

5.2.3 Etching

Pattern transfer uses a resist layer as a etchant mask to allow selective removal of the underlying material. The process of etching is critical to creating fine-line features on integrated circuits as it establishes the limits on the pattern resolution.

The basic problem in etching is illustrated by the example in Fig. 5.21(a), where a mask (resist) layer has been patterned to allow etching of an oxide. The etch process is characterized by the *etch rates* v_{ver} and v_{lat}, which represent the rate of etching in the vertical and lateral directions, respectively. Practical units for etch rates are [Å/min]. The two-dimensional nature of the etching process is specified by the *degree of anisotropy*

$$A = 1 - \frac{v_{lat}}{v_{ver}}, \tag{5.2-1}$$

where A varies from 0 to 1 depending on the ratio of the etch rates.

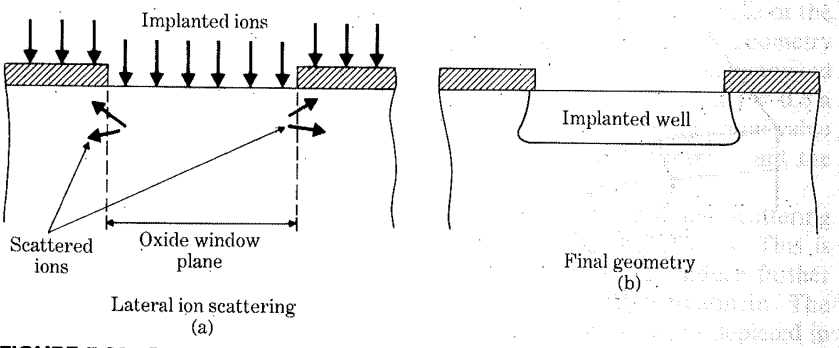


FIGURE 5.20 Lateral doping effects in an ion implantation.

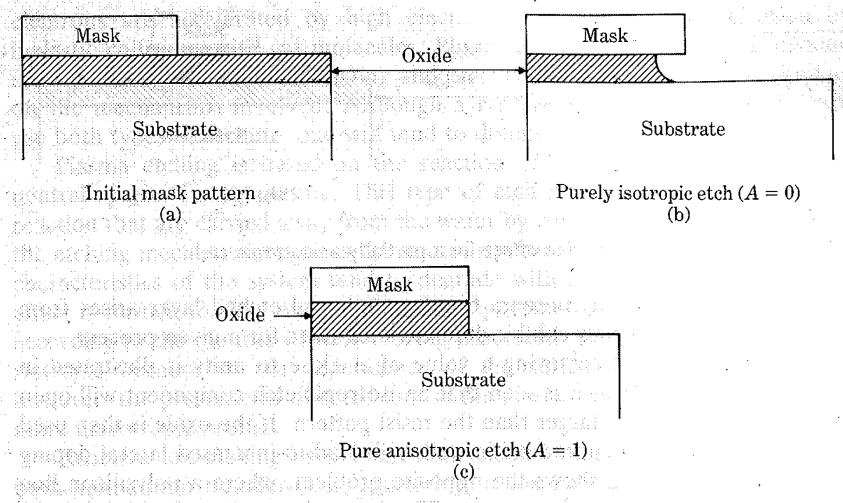


FIGURE 5.21 Basic etching problem.

Purely isotropic etching occurs when $v_{lat} = v_{ver}$, i.e., $A = 0$. This results in the structure shown in Fig. 5.21(b), where the oxide etch undercuts the resist mask. A *totally anisotropic* etch is defined by $v_{lat} = 0$, giving $A = 1$. In this case, only vertical etching is present so that the oxide wall is perpendicular to the substrate surface as shown in Fig. 5.21(c). Anisotropic etching thus results in the ideal structure.

Figure 5.22 shows the general case where a film of thickness h is etched with an amount ℓ lost in the lateral undercut. This is described by

$$A = 1 - \frac{\ell}{h} \tag{5.2-2}$$

since the amount of etching is just the etch rate multiplied by the etch time t_e. The drawing also gives a more realistic case in which the masking material itself has undergone some etching. This leaves a slight tilt in the mask edge.

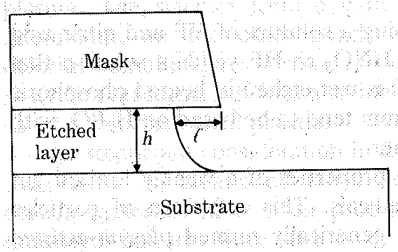


FIGURE 5.22 Geometry for a partially anisotropic etch.

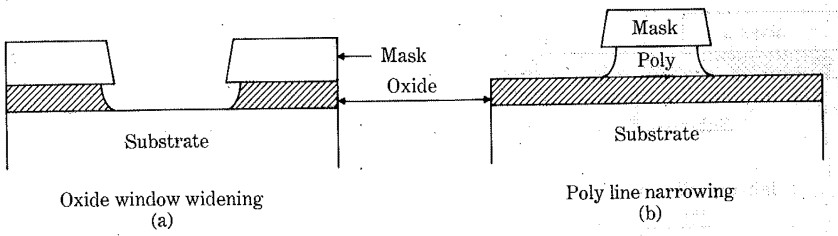


FIGURE 5.23 Window and line effects in a partially anisotropic etch.

The difference in etched contours for the mask and etched layer arises from the fact that two materials exhibit different etch rates for a given process.

The importance of achieving a value of A close to unity is illustrated in Fig. 5.23. In Fig. 5.23(a) it is seen that an isotropic etch component will open an oxide window that is larger than the resist pattern. If the oxide is then used as a dopant mask, the curved oxide walls will lead to increased lateral doping spreads. Figure 5.23(b) shows the opposite problem, where a polysilicon line feature is defined by a resist pattern. The isotropic portion of the poly etch from v_{lat} gives a linewidth smaller than the resist pattern as shown by the cantilever resist structure. This decreases the cross-sectional area of the current flow, which increases the sheet resistance of the line. In addition, the undercutting requires that the mask pattern be larger than the final chip feature.

Etching techniques are generally classified as being either *wet* or *dry*. *Wet etching,* as implied by its name, employs aqueous solutions of acids (or caustics) to etch the desired layers. The wafers are immersed directly in the solution, which allows the chemical reaction to take place over the exposed surface regions. Wet etches tend to be isotropic when used on amorphous or polycrystalline material. This then limits the use of wet etching to lithographic features larger than about 3 [μm].

Silicon dioxide may be wet-etched in a buffered hydrofluoric acid solution containing HF, NH_4F, and H_2O. The HF readily attacks the SiO_2 layer but does not etch bare silicon. This property of *selectivity* makes it useful for etching oxide windows down to a silicon substrate since the vertical etch rate goes to zero when the substrate is reached. The HF solution is buffered to slow down the etch rate to a controllable level.

Phosphosilicate glass can be etched using a solution of HF and nitric acid (HNO_3) in water. Increasing the ratio of HNO_3 to HF yields a solution that will etch polysilicon. Silicon nitride (Si_3N_4) is wet-etched in heated phosphoric acid (H_3PO_4). Solutions for etching aluminum tend to be based on H_3PO_4 with HNO_3 and CH_3COOH (acetic acid) in water.

Dry etching systems are based on the properties of a weakly ionized gas consisting of electrons, ions, and free radicals. This collection of particles constitutes a *plasma,* so this approach is generically termed *plasma-assisted* etching. The plasma is usually produced by an rf glow discharge in which

electrons are accelerated by high electric fields and induce ionization by inelastic collisions with gas molecules. Plasma-assisted etching can be divided into two categories, *plasma etching* and *reactive ion etching* (RIE), depending on the mechanisms involved. Although a realistic dry etch arrangement will use both types of etching, one will tend to dominate the overall process.

Plasma etching is based on the reaction of the surface material with a neutral species in the plasma. This type of etch produces by-products of the reaction that are carried away from the wafer by controlling the gas flow. Since the etching mechanism requires the neutral species in the reaction, the etching characteristics of the system tend to degrade with increased wafer area. This effect is termed *loading* and requires an adjustment of the system parameters according to the number of wafers being processed. Straight plasma etches are isotropic because of the dependence of the surface reaction. In practice, however, the presence of directed ions in the plasma can enhance v_{ver} to give a more anisotropic profile.

Reactive ion etching uses a directed ion flux that etches the material by both momentum transfer (i.e., sputtering or *ion milling*) and a reactive etch mechanism. RIE systems are generally designed so that the ions impact the wafer vertically to ensure that $v_{ver} \gg v_{lat}$. This directionality allows for a minimization of the lateral etching, so that highly anisotropic etch contours are possible. RIE is thus a good choice for fine-line VLSI structures.

The gases used to create the plasma determine both the etch rates and profiles for a given layer. Polysilicon and silicon compounds (SiO_2, Si_3N_4, PSG, etc.) can be etched using carbon tetrafluoride (CF_4). The basic plasma forming reaction is

$$CF_4 + e^- \rightarrow CF_3^+ + F^* + 2e^-,$$

which shows how an electron (e^-) impacts with a CF_4 molecule to produce the CF_3^+ ion and the F* free radical. The F* radical is the active species for reactively etching silicon by forming SiF_4, which is stable and can be removed from the wafer surface. A similar situation holds for silicon nitride etching. Silicon dioxide relies on the CF_3^+ ion for intermediate reduction to SiO, which then reacts to form SiF_4. It is found that addition of oxygen (and other gases) can greatly affect the etching characteristics of the plasma for certain materials.

The refractory metals Mo, W, Ti, and Ta can be etched using a CF_4 plasma. The silicides formed with these metals also exhibit this property. Aluminum, on the other hand, cannot be etched with CF_4. Rather, a chlorine-containing gas such as carbon tetrachloride (CCl_4) is used to induce a reaction that forms $AlCl_3$ as a stable by-product. Polysilicon can also be etched using a chlorine-based plasma.

An important consideration in dry etching centers around the selectivity of the plasma. As mentioned above, selectivity refers to the ability to etch only certain materials in a given process. This is particularly crucial in etching multilayer patterns since each layer reacts differently to the etch mechanisms. In wet etching, selectivity is attained by using acids that attack only a single

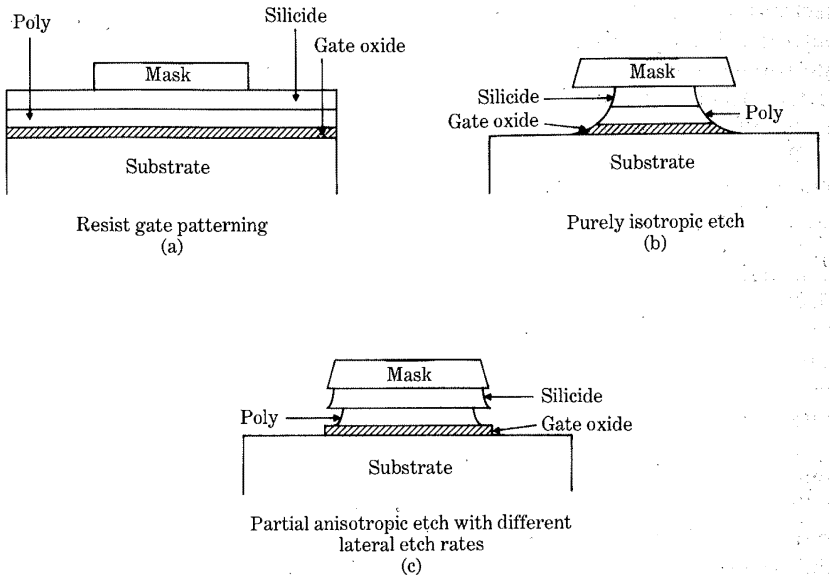


FIGURE 5.24 Etch selectivity problem for a polycide MOSFET gate.

layer of material at a time. This can be partially achieved in a purely reactive dry etch by varying the gas composition to preferentially etch a given layer at a faster rate. Note, however, that a plasma made with CF_4 can potentially etch any silicon-containing layer, so that high selectivity is difficult to obtain. Another problem in dry etching is that ion bombardment of the wafer sputters off material at a rate that is relatively insensitive to the composition. Consequently, there are no built-in mechanisms to stop the process after the layer is etched. This must be controlled externally by monitoring the etching time to prevent over-etching the wafer.

An interesting example of the multilayer dry etching problem is provided by the polycide-gate MOSFET in Fig. 5.15 of the previous section. Patterning of the gate requires etching through the silicide, the polysilicon, and the gate oxide. Figure 5.24(a) shows the initial layering with the gate etch mask. If a purely isotropic plasma etch is used, the gate will have the structure shown in Fig. 5.24(b). Although this gives a smooth transition at the layer interfaces, there is significant undercutting of the silicide underneath the mask. This results in increased sheet resistances, which act against the basic reason for introducing the silicide layer in the first place. Figure 5.24(c) shows the structure obtained using an RIE in which the lateral etch rate for the poly layer is greater than that for the silicide. This produces an undesirable silicide cantilever in the MOSFET. Problems of this type illustrate the importance of characterizing the etching process used in the wafer fabrication.

5.3 Isolation Techniques

Isolation in an integrated circuit deals with the need to electrically isolate neighboring devices on the chip. MOSFETs are fabricated on planar portions of the silicon known as *active areas*. Active areas are surrounded by field regions, which are included for the isolation. Interconnects are patterned conducting layers for "wiring" the devices together according to the circuit design. Isolation sets the characteristics of the field regions to prevent unwanted conduction paths between active areas. This, of course, is required to ensure that the circuit operates properly.

A MOS technology is partially self-isolating in that the MOSFETs have *pn* junctions that are usually reverse-biased. For example, an *n*-channel MOSFET consists of n^+ drain and source wells in a *p*-type substrate. Since only leakage current flows across the reverse biased junctions, some isolation is automatically provided when the transistors are placed into the chip environment.

The main complicating factor in MOS isolation is that unwanted inversion layers may form under the interconnect lines that are patterned in the field regions. If the field silicon surface becomes inverted, a conduction path may be established between MOSFETs in adjacent active areas. Isolation techniques are usually directed toward dealing with this problem.

The basic requirements for an isolation scheme are that (a) the technique should eliminate the possibility of forming unwanted field inversion layers and (b) the initial active area regions must be flat for fabricating the MOSFETs. Another desirable feature is that the overall chip process yield planar surface topologies after each major step. There are different approaches used for isolation in MOS integrated circuits. Three are presented below.

5.3.1 Etched Field Oxide Isolation

This basic MOS isolation approach uses the sequence shown in Fig. 5.25 to define and isolate active areas. The starting point is a lightly doped *p*-type

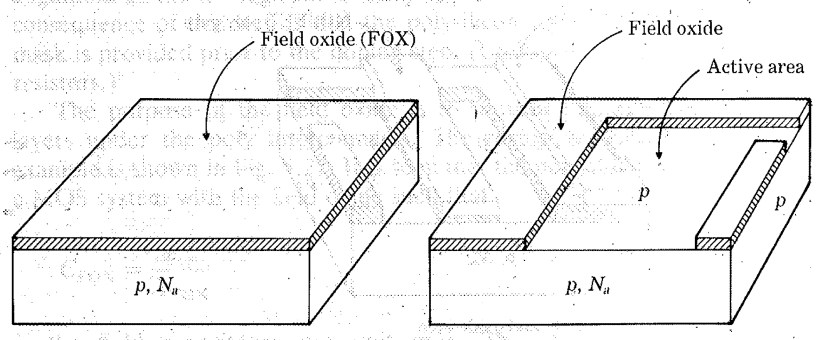


FIGURE 5.25 Etched field oxide isolation.

wafer that serves as the substrate. The first step is to thermally grow a thick *field oxide* (FOX) over the entire surface of the wafer, as illustrated by Fig. 5.25(a). Active areas are defined by a masking step that selectively etches the field oxide down to the bare silicon. The patterning for the present example is shown in Fig. 5.25(b). The field is defined by regions of the wafer where the FOX remains after this etch step. Active area patterns are set by the surface geometry requirements for the MOSFET to be placed in each location. This process also allows for n^+ line patterns to be formed during the active area definitions. These serve as silicon-level interconnects.

The isolation characteristics of the field oxide are best understood by examining the structure of a completed MOS chip section. A poly gate enhancement-mode MOSFET can be fabricated using the sequence of Fig. 5.26. The first step is the growth of the thin MOSFET gate oxide with thickness x_{ox} as illustrated in Fig. 5.26(a). Next a p-type ion implant (e.g., BF_2) is performed to induce a positive threshold voltage shift for an E-mode

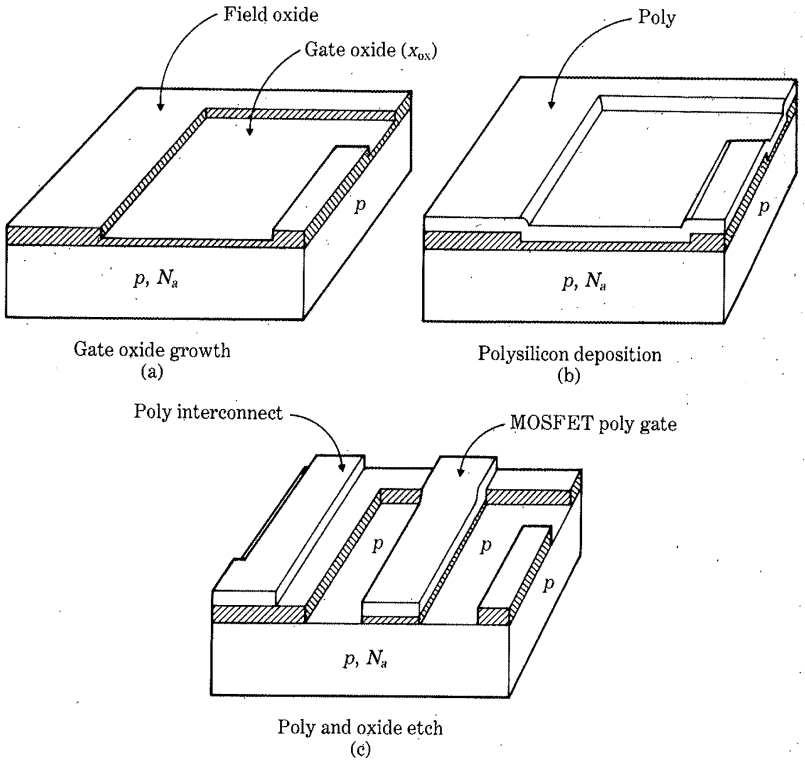


FIGURE 5.26 Enhancement-mode MOSFET fabrication sequence for etched field oxide isolation.

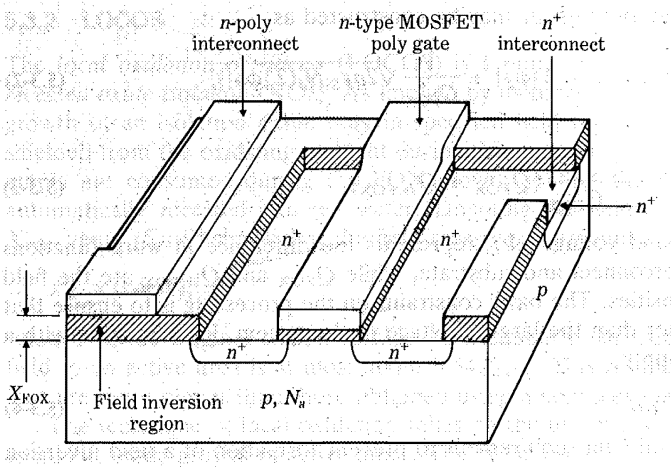


FIGURE 5.27 Final enhancement-mode MOSFET in an etched field oxide isolation process.

transistor. The gate polysilicon layer is then deposited over the entire surface, giving the layering shown by Fig. 5.26(b). The poly layer is also used as an interconnect level. The poly patterning is obtained from a masking step, and then the gate oxide is etched to leave the basic structure of Fig. 5.26(c). This shows a MOSFET pattern in the active area with a poly interconnect line running over the adjacent field region.

The final step needed to complete the basic MOSFET is to perform an n-type doping with either a diffusion or an ion implantation. This results in the completed MOSFET-interconnect arrangement in Fig. 5.27. The drain and source n^+ wells have been patterned using the gate (and FOX) as a mask. This is called a *self-aligned* MOSFET since the gate pattern allows for automatic alignment of the n^+ regions. Overlap capacitance is thus minimized. A direct consequence of this step is that the polysilicon layer is doped n-type unless a mask is provided prior to the doping step. (Undoped poly lines may be used as resistors.)

The purpose of the field oxide is to prevent the formation of inversion layers under the poly interconnects. The region of concern in the present example is shown in Fig. 5.27. It is seen that the polysilicon interconnect forms a MOS system with the field oxide such that

$$C_{FOX} = \frac{\varepsilon_{ox}}{X_{FOX}} \tag{5.3-1}$$

is the field capacitance per unit area. The process flow is set to give $X_{FOX} \gg x_{ox}$. Consequently, $C_{FOX} \ll C_{ox}$, indicating a reduced level of MOS field coupling into the substrate.

The *field threshold voltage* may be constructed as

$$(V_{T0})_{\text{FOX}} = V_{FB,\text{FOX}} + 2|\phi_F| + \frac{1}{C_{\text{FOX}}}\sqrt{2q\varepsilon_{\text{Si}}N_a(2|\phi_F|)}, \tag{5.3-2}$$

where

$$V_{FB,\text{FOX}} = \Phi_{IS} - \frac{1}{C_{\text{FOX}}}(Q_{\text{FOX}} + Q_{ss,\text{FOX}}) \tag{5.3-3}$$

is the field flatband voltage. Φ_{IS} represents the difference in work functions between the interconnect and substrate, while Q_{FOX} and $Q_{ss,\text{FOX}}$ are the field oxide charge densities. The basic constraint on the processing is to ensure that $(V_{T0})_{\text{FOX}}$ is greater than the largest voltage in the system. For example, with a single voltage supply,

$$(V_{T0})_{\text{FOX}} > V_{DD} \tag{5.3-4}$$

represents the minimum requirement to prevent formation of a field inversion layer.

The value of the threshold voltage may be increased by two techniques. The first is to provide a field acceptor implant with a dose $D_{I,\text{FOX}}$ that increases the field threshold voltage by an amount

$$(\Delta V_{T0})_{\text{FOX}} = \frac{qD_{I,\text{FOX}}}{C_{\text{FOX}}}. \tag{5.3-5}$$

The field implant must precede the initial field oxide growth. The resulting p^+ under the FOX is sometimes referred to as a *channel stop.*

An alternate approach is to apply a body bias voltage V_B to the p-type substrate. Assuming that V_B is negative with respect to ground, this induces a body bias threshold voltage shift of

$$(\Delta V_T)_{\text{FOX}} = \gamma_{\text{FOX}}(\sqrt{2|\phi_F| + |V_B|} - \sqrt{2|\phi_F|}), \tag{5.3-6}$$

where

$$\gamma_{\text{FOX}} = \frac{1}{C_{\text{FOX}}}\sqrt{2q\varepsilon_{\text{Si}}N_a} \tag{5.3-7}$$

is the field body bias factor. Using this technique requires that V_B be accounted for in the circuit design since all MOSFET threshold voltages will increase. It should be noted that the threshold voltage shifts described in the above formulas are proportional to X_{FOX}.

The main drawback of etched field oxide isolation is that $X_{\text{FOX}} \gg x_{\text{ox}}$ implies the existence of large oxide "steps" at the boundaries between active areas and field regions. Transitions of this type can lead to problems in depositing a conformal (uniform) layer of poly over the isolation terrain. For example, a crack may occur in the poly layer that could lead to chip failure. This gives motivation for examining isolation techniques that recess the field oxide into the silicon, yielding a more planar overall working surface.

5.3.2 LOCOS

The *local oxidation of silicon* (LOCOS) is a popular approach for achieving *recessed oxide isolation* (ROI). As implied by its name, LOCOS allows for the growth of an isolation oxide only in specified field regions. Active areas are shielded from the oxidizing ambient during this processing step. Surface silicon atoms are consumed during the FOX growth so that the isolation oxide is automatically recessed into the wafer. Denoting the field oxide thickness by X_{FOX}, eqn. (5.1-9) shows that the field recesses an amount

$$x_{Si} \simeq 0.46 X_{FOX} \tag{5.3-8}$$

relative to the original silicon surface. Consequently, an oxide step from the field to an active area is at most about $0.54X_{FOX}$. This results in chip surfaces that are more planar than those obtained using a nonrecessed isolation oxide.

The technique of local oxidation relies on the use of a silicon nitride layer to inhibit oxide growth in active areas. The basic sequence is illustrated in Fig. 5.28. First, a stress-relief oxide layer and a nitride layer are patterned to define the desired active area geometry, shown in Fig. 5.28(a). Next an acceptor field implant is performed to increase the value of $(V_{T0})_{FOX}$. The resulting p^+ regions are shown in Fig. 5.28(b). The field implants are self-aligned to the active areas by the nitride masking.

Localized growth of the field oxide is achieved using the fact that Si_3N_4 does not react with O_2 and only slowly oxidizes in steam. A typical FOX growth might start with an initial growth in chlorinated O_2, followed by a long steam oxidation that produces most of the field oxide. The wafer would then be subjected to a dry O_2 anneal/growth step to improve the quality of the steam-grown oxide. The field regions (not covered by the nitride) are directly oxidized by this sequence of steps. The resulting structure appears as shown in Fig. 5.28(c).

The "lifting" of the nitride edges is easily understood. During the FOX growth, some of the oxidizing gas will diffuse through the stress-relief oxide to the underlying silicon. This induces an oxide growth around the edges of the nitride pattern, which in turn lifts the nitride layer. The transition from the stress-relief oxide to the FOX is called the *bird's beak* region because of its shape. The formation of the bird's beak in LOCOS results in a smaller active (planar) area than that originally defined by the nitride patterning. The loss of active area to the bird's beak is termed *encroachment.* The amount of encroachment increases with increasing thickness of the stress-relief oxide layer, since more oxygen can diffuse under the nitride pad.

MOSFETs are made in the active areas by first removing the nitride (Fig. 5.28d) and then etching the stress oxide down to the silicon (Fig. 5.28e). This allows for a controlled growth of the critical gate oxide with thickness x_{ox}, as illustrated by Fig. 5.28(f). The working value of the field oxide thickness X_{FOX} is also established at this point since the next layering will be the polysilicon gate.

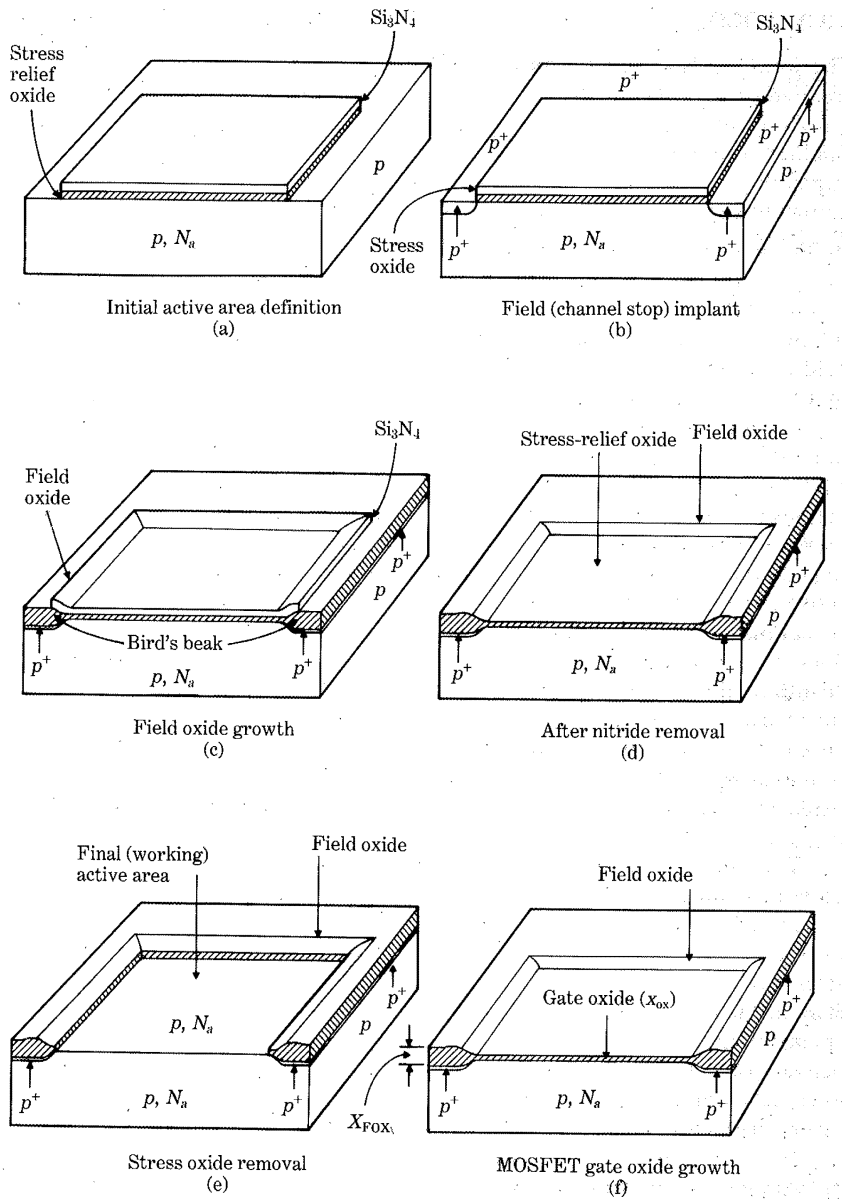


FIGURE 5.28 Basic LOCOS MOS isolation process flow.

LOCOS is a relatively complicated process to characterize and generally requires the use of numerical simulation studies. Although the details of the analysis are beyond the scope of the present treatment, three points deserve further comment.

First, it should be noted that the acceptor field implant used to increase $(V_{T0})_{FOX}$ must be relatively deep with a moderate dose level. This arises because the field oxide growth will consume the surface portions of the doped p^+ layer. The initial implant characteristics must be adjusted to account for this part of the processing. The dose level of the implant must be high enough to prevent the field from inverting. However, increasing the field implant density increases the MOSFET n^+/substrate depletion sidewall capacitance C_{j0sw}, as shown by eqn. (4.6-21). This consideration requires that only moderate implant doses (10^{12}–10^{13} [cm^{-2}]) be used.

The second point is concerned with the reaction of silicon nitride and steam by means of

$$Si_3N_4 + 6H_2O \rightarrow 3SiO_2 + 4NH_3.$$

The ammonia (NH_3) produced by this reaction tends to diffuse through the structure and can eventually make it to the underlying silicon. The ammonia will then react to form a layer of Si_3N_4 at the silicon surface. This is most prevalent around the edge of the nitride pattern and gives rise to what is sometimes called the "white ribbon" effect. The presence of this nitride can cause nonuniform gate oxide growths, so it must be removed with an additional processing step.

The last comment deals with the formation of the bird's beak and the subsequent active area encroachment. This is an important factor in chip layout since it requires that the initial active area definition be larger than the final working value. A typical order of magnitude might be 0.8 [μm] lateral encroachment (per side) using a 200 [Å] stress-relief oxide thickness; the actual value varies according to the process specifications. Encroachment is particularly crucial for VLSI circuits since it creates a fundamental limitation on high-density layouts.

5.3.3 Trench Isolation

LOCOS is a widely used technique since it provides the necessary isolation properties and is relatively straightforward to implement. However, the undesirable active area encroachment associated with the formation of the bird's beak has spurred development of new isolation schemes for use in high-density chip layouts. This section will discuss the use of "trenches" as an alternate approach to LOCOS.

Trench isolation uses a highly anisotropic silicon etch to create vertical-walled trenches around active areas. The trenches are filled with a dielectric to establish the boundaries needed to ensure electrical isolation. The process has the advantage that it exhibits only minor bird's beaking and negligible

encroachment. In addition, trenches provide isolation structures that are almost completely recessed into the wafer.

The basic process flow involved in a trench isolated integrated circuit is illustrated in Fig. 5.29. Figure 5.29(a) shows a stress-relief oxide with a nitride layer used to define the active areas. The example chosen here also uses a CVD oxide layer as a mask in the trench patterning. Once the active areas are defined, the wafer is subjected to an anisotropic silicon etch (usually an RIE) that gives the initial formation of the trench structuring shown in Fig. 5.29(b). At this point, boron is implanted to create a p^+ channel-stop layer at the bottom of the trench. The oxide mask is then removed.

Ideally, isolation would be accomplished by filling the trench with an insulating dielectric such as silicon dioxide. Unfortunately, it is difficult to fill the trench using a CVD oxide because of void formation and other problems in trying to obtain uniform coverage. To overcome this, the wafer is first oxidized to grow oxide layers on the bottom and sidewalls; this is shown in Fig. 5.29(c). Then polysilicon is deposited over the structure, filling the trenches. The surface poly is etched, leaving the filled trenches illustrated in Fig. 5.29(d). The last step needed in the isolation process flow is to oxidize the wafer. This gives the final cross-section in Fig. 5.29(e).

This typical sequence illustrates some important points. First, note that there is a small amount of active area encroachment during the thermal oxidation of the trench walls in Fig. 5.29(c). Additional encroachment may occur during the poly oxide growth, which gives the surface oxide layer of Fig. 5.29(e) due to both silicon consumption and any bird's beak formation under the nitride. The magnitude of the encroachment is relatively small (particularly when compared with a standard LOCOS process), which makes this type of isolation a good candidate for high-density circuit layouts.

The use of polysilicon as a "filler" material for the trench provides good coverage but can introduce complications in the isolation characteristics. The basic problem is shown in Fig. 5.30(a), where an interconnect runs over the trench. Owing to normal coupling capacitances, the polysilicon inside the trench acts as an electrode with a floating potential.

There are three values of oxide thickness that enter into the analysis: the top oxide X_{TOX}, which is grown during the final oxidation step; the sidewall oxide X_{SOX}; and the bottom oxide X_{BOX}. Although the thickness of the trench sidewall and bottom oxides are usually different (due, for example, to nonuniform surface oxidation rates), $X_{BOX} = X_{SOX}$ will be assumed for simplicity. The important capacitances per unit area are thus given by

$$C_{TOX} = \frac{\varepsilon_{ox}}{X_{TOX}} [\text{F/cm}^2]$$

$$C_{SOX} = \frac{\varepsilon_{ox}}{X_{SOX}} = C_{BOX} [\text{F/cm}^2]. \tag{5.3-9}$$

At the present time, typical values for the oxide layers are $X_{SOX} = 1500$ [Å] and $X_{TOX} = 4000$ [Å].

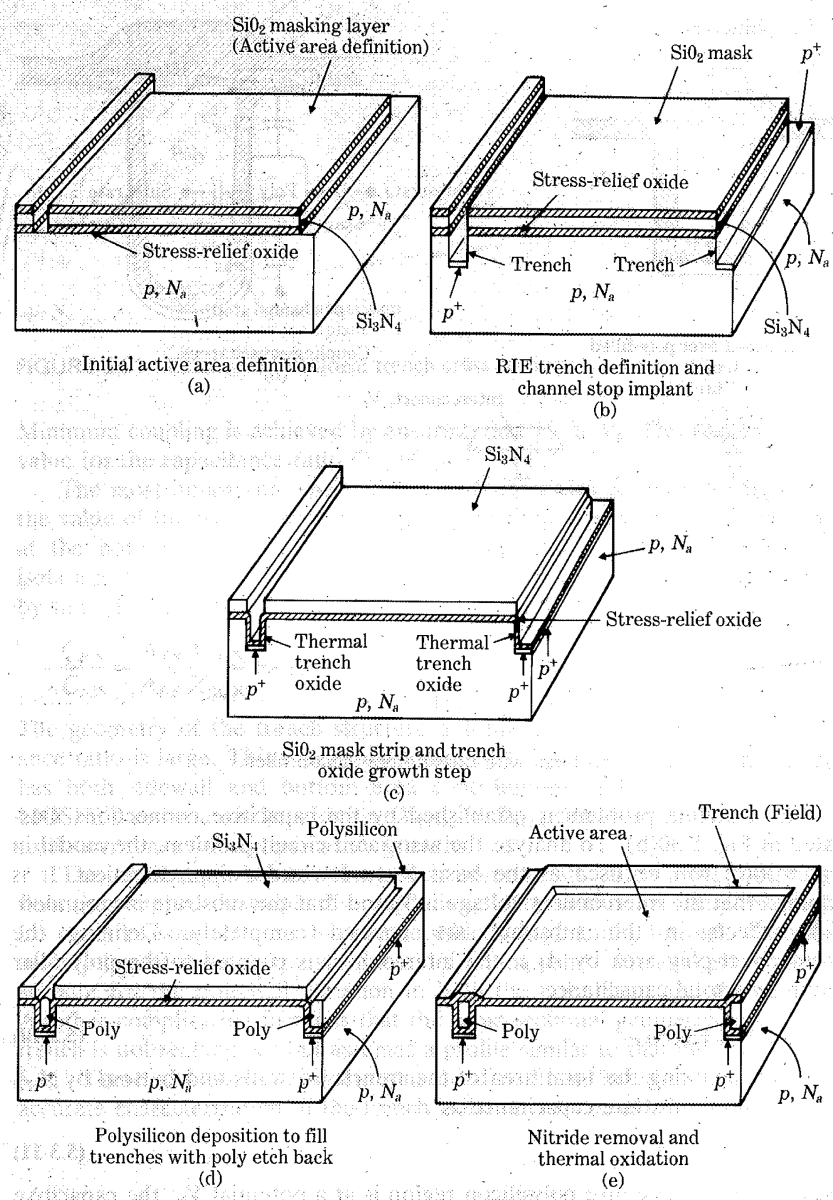


FIGURE 5.29 Basic sequence for trench isolated integrated circuits.

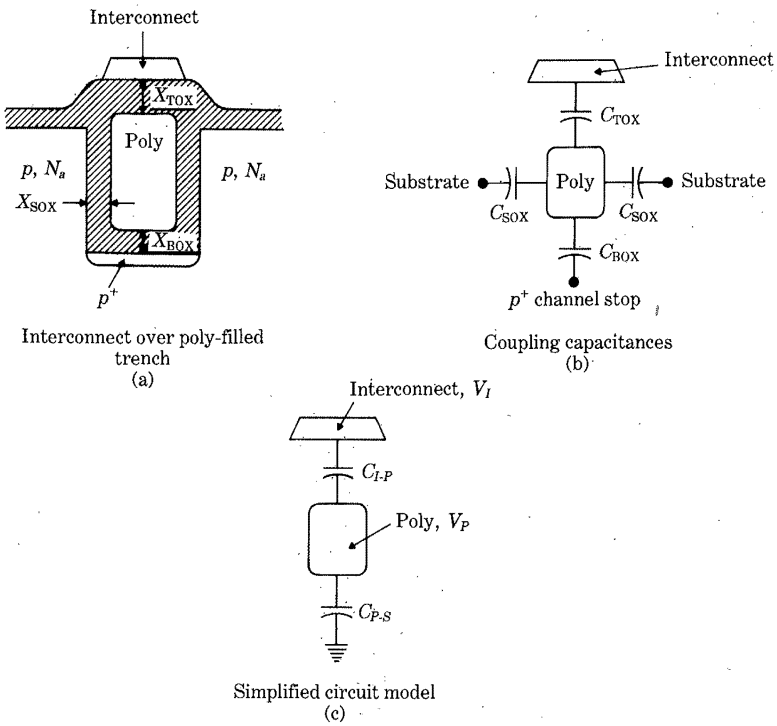


FIGURE 5.30 Coupling problem with floating polysilicon filler.

The coupling problem is established by the capacitive connections illustrated in Fig. 5.30(b). To analyze the associated circuit problem, the model in Fig. 5.30(c) will be used as the basis for a low-order approximation. It is assumed that the interconnect voltage is V_I and that the substrate is grounded. Field effects in the substrate are ignored completely. Defining the interconnect-poly area by $A_{I\text{-}P}$, the interconnect is coupled to the poly filler region by a total capacitance

$$C_{I\text{-}P} = C_{TOX}A_{I\text{-}P}. \tag{5.3-10}$$

Similarly, denoting the total area of the trench sidewalls and bottom by A_{TR} gives the poly-substrate capacitance as

$$C_{P\text{-}S} = C_{SOX}A_{TR}. \tag{5.3-11}$$

Assuming that the entire polysilicon region is at a potential V_P, the capacitive voltage divider circuit gives

$$V_P = \frac{C_{I\text{-}P}}{C_{I\text{-}P} + C_{P\text{-}S}} V_I. \tag{5.3-12}$$

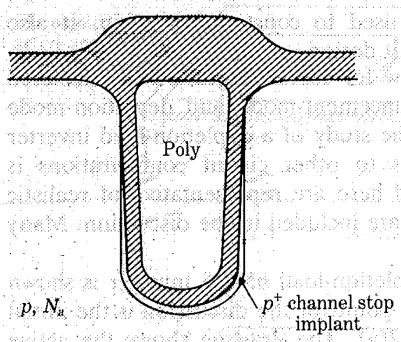


FIGURE 5.31 Example of nonideal trench cross section.

Minimum coupling is achieved by ensuring that $V_P \ll V_I$. This requires a large value for the capacitance ratio $C_{P\text{-}S}/C_{I\text{-}P}$.

The most important aspect of this analysis is that V_P must be kept below the value of the trench threshold voltage $(V_{T0})_{BOX}$ needed to invert the silicon at the bottom of the trench. This, of course, is required to preserve the isolation characteristics. Equation (5.3-12) shows that $V_P \ll V_I$ can be attained by satisfying the condition

$$\frac{C_{P\text{-}S}}{C_{I\text{-}P}} = \frac{A_{TR}}{A_{I\text{-}P}} \frac{X_{TOX}}{X_{SOX}} \gg 1. \tag{5.3-13}$$

The geometry of the trench structure automatically ensures that the capacitance ratio is large. This is seen by noting that $A_{TR} \gg A_{I\text{-}P}$ because the trench has both sidewall and bottom area contributions and surrounds an entire MOSFET active area. In addition, $X_{TOX} > X_{SOX}$ can be set by controlling the oxidation times for the two oxide growths.

The problems involved with trench isolation are much more complicated than the simplistic example discussed above. This additional complexity arises for many reasons. One is that the polysilicon is generally intrinsic so that the assumption of a single poly voltage V_P is not valid. An analysis must then account for the potential variation in both the polysilicon and the substrate. Another complicating factor is that the cross-sectional geometry of a realistic trench is not rectangular but assumes a profile similar to that illustrated in Fig. 5.31. These and other considerations need to be included to obtain a more accurate characterization of the trench isolation structures.

5.4 nMOS LOCOS Process Flow

A *process flow* describes the sequence of fabrication steps used to create an integrated circuit. Knowledge of the process flow allows for a complete

electrical characterization of the layers used to construct the chip. It also provides the information needed for layout design.

This section deals with a process flow for fabricating a LOCOS-isolated nMOS integrated circuit with both enhancement-mode and depletion-mode transistors. Emphasis will be placed on the study of a depletion-load inverter on the chip. Extension of the concepts to other circuit configurations is straightforward. The sequences presented here are representative of realistic processes. However, only the basic steps are included in the discussion. Many variations are possible.

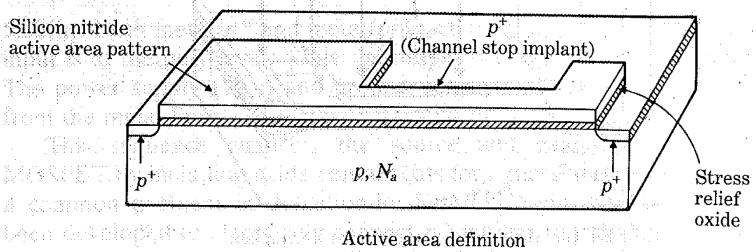
The process flow for fabricating a depletion-load nMOS inverter is shown by the drawings in Fig. 5.32. The starting point for the discussion is the initial active area patterning shown in Fig. 5.32(a). The drawing shows the active area geometry defined using the stress-relief oxide and nitride layers. In addition, a self-aligned p^+ channel stop field implant has been performed using the nitride patterning. For future reference it is noted that the E-mode driver MOSFET will be on the left side of the active area, while the D-mode load transistor will be placed on the right side.

The field oxide (FOX) used in the LOCOS isolation scheme is grown by subjecting the wafer to an oxidizing ambient. This results in the semirecessed field oxide structure shown in Fig. 5.32(b). Note the characteristic bird's beak formation around the nitride edges. The next steps are to remove the nitride and then etch the stress-relief oxide, which gives access to the silicon surface as in Fig. 5.32. Once this is accomplished, the critical MOSFET gate oxide with thickness x_{ox} is grown. Figure 5.32(d) shows the structure at this point in the process flow.

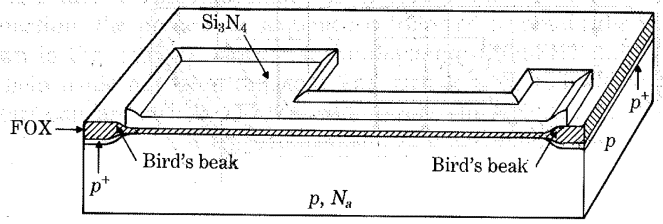
Two threshold adjustment ion implant steps are included in the fabrication sequence. The first is an acceptor implant used to set the threshold voltage $V_{T0} > 0$ of the enhancement-mode MOSFETs. Resist is used as a mask to selectively implant the E-mode active area; the patterning for this step is shown in Fig. 5.32(e). Implanted p-type threshold adjustment regions are not usually shown explicitly in cross-sectional drawings. Depletion-mode MOSFETs are made using the donor implant in Fig. 5.32(f). This step creates a patterned n-well, which eventually serves as the D-mode transistor channel. The use of donors ensures that the threshold voltage satisfies $V_{T0} < 0$.

Polysilicon is deposited over the gate oxide in the next step. The poly layer is patterned, and the gate oxide is etched. An n^+ implant is then performed to create self-aligned drain and source regions. The structure now appears as in Fig. 5.32(g), where the identities of the transistors become clear. Note that the n^+ implant uses the polysilicon layer as a mask, so the poly layer is doped heavily n-type unless a separate mask is provided to keep the layer undoped. This is important for computing both threshold voltages (in Φ_{GS}) and parasitic line resistances.

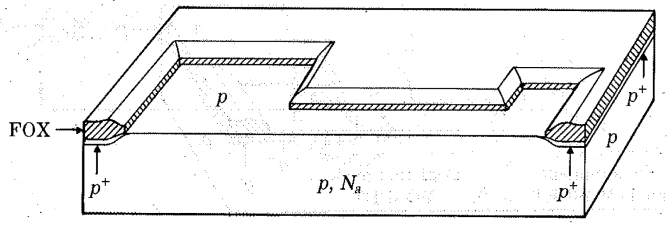
The last steps in the basic processing sequence are to coat the surface with CVD oxide and then pattern contact cuts where needed. Metal is deposited over the surface and patterned. The final inverter structure is shown in Fig.



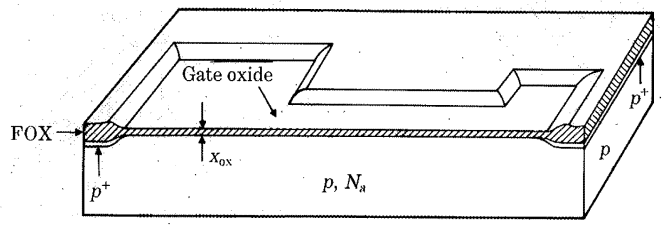
Active area definition
(a)



LOCOS field oxide (FOX)
isolation growth
(b)

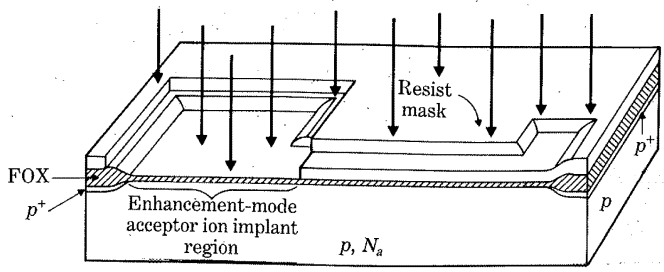


After nitride and stress-relief oxide
removal
(c)

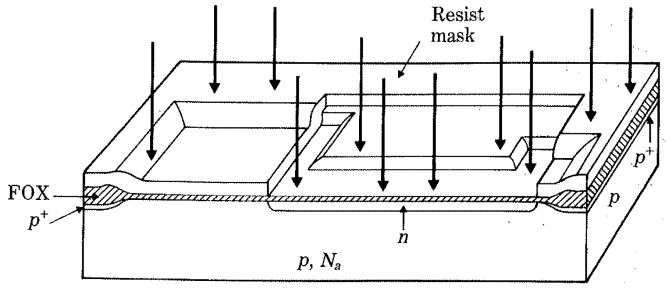


Gate oxide growth
(d)

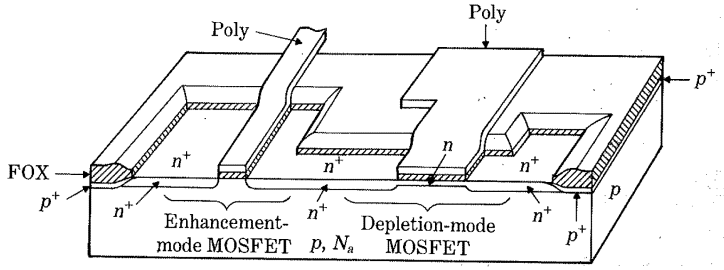
FIGURE 5.32 Basic nMOS LOCOS process flow for depletion-load inverter.



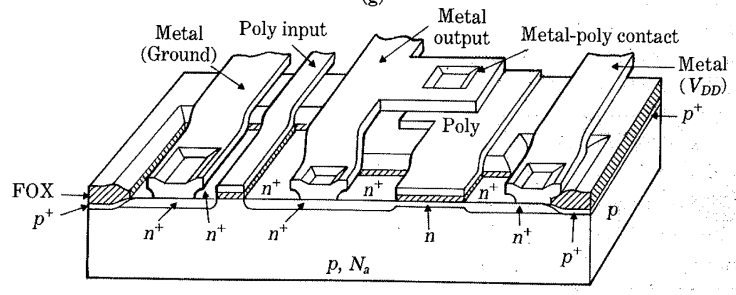
E-mode MOSFET threshold voltage adjustment ion implant
(e)



Depletion-mode MOSFET donor ion implant
(f)



Poly deposition/patterning, oxide etch and n^+ implant
(g)



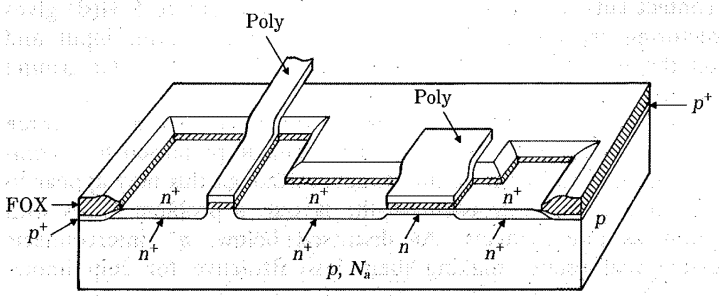
Completed inverter after metallization
(h)

Figure 5.32 (*contd.*)

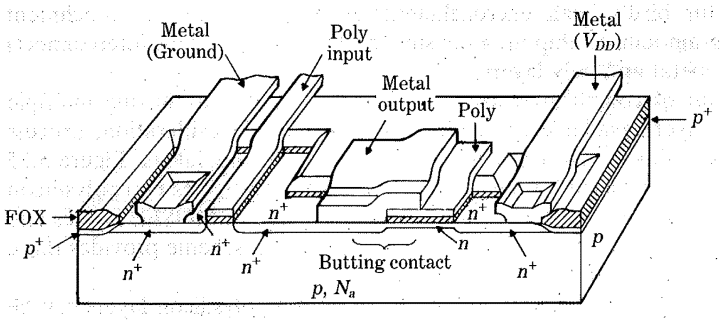
5.32(h). Both metal/n^+ and metal/poly contacts have been made. The inverter input is at the poly level, while the output is chosen to be on the metal level. The power supply (V_{DD}) and ground connections to the circuit are obtained from the metal lines.

This sequence connects the source and gate of the depletion-load MOSFET by including oxide contact cuts for a metal interconnect. Since this is a common problem in depletion-load nMOS logic, alternate schemes have been developed to electrically connect n^+ regions to poly layers. Two of these techniques are included in the discussion here. In general, both will result in reduced real estate consumption.

The first n^+/poly technique is termed a *butting contact.* To effect this connection, the processing sequence is followed to obtain the poly patterning shown in Fig. 5.33(a). Note that the depletion MOSFET poly gate area over the field oxide has been reduced. The surface is coated with oxide, and then contact cuts are etched into the oxide layer. The structure of a butting contact



After poly patterning and n^+ implant
(a)



After metallization
(b)

FIGURE 5.33 Formation of a butting metal/poly/n^+ contact.

is illustrated in Fig. 5.33(b). Metal has been deposited such that it directly connects the poly gate to the n^+ region. The savings in real estate is evident from the metal/n^+ and n^+/poly contacts having merged into one. The main drawback of a butting contact is that the metal may crack because of the transition from the top of the poly to the n^+ layer.

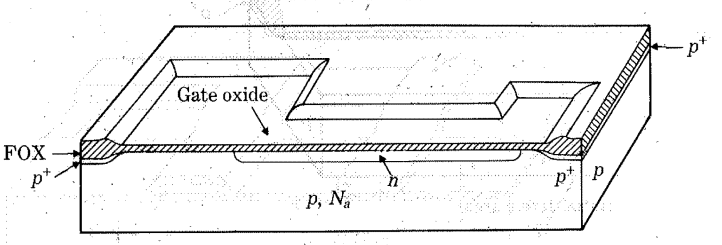
Buried contacts achieve n^+/poly connections without using the metal layer. A typical sequence is illustrated in Fig. 5.34. The starting point for creating a buried contact is directly after the threshold adjustment implants have been performed. This is shown as Fig. 5.34(a). A buried contact relies on the fact that an n-type layer has been created for D-mode MOSFETs. The first step in creating the buried contact is to perform a selective etch into the gate oxide as in Fig. 5.34(b). Next the polysilicon is deposited and patterned so that the depletion MOSFET gate poly extends over the exposed n-region. Figure 5.34(c) gives the polysilicon patterning. The n^+ drain-source implant establishes the buried contact, since the polysilicon and n^+-wells are now electrically connected through the n-region. The final steps are to coat the wafer with oxide, etch contact cuts, and perform the metallization. Figure 5.34(d) gives the final structuring chosen for the example. Both the inverter input and output are on the poly level. The metal is used only for V_{DD} and ground connections in the gate.

The process flows described so far are characterized by having three interconnect layers; metal, poly, and n^+. These can be patterned to accommodate the wiring dictated by the circuit design. Although this may appear to allow a significant amount of freedom in the layout, a problem arises with using n^+-regions as interconnects. As discussed below, n^+ interconnects require excessive real estate, making them less attractive for chip implementations.

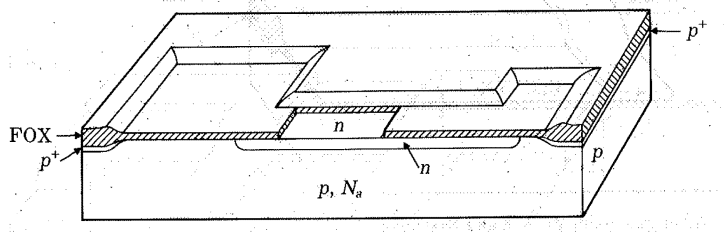
In the LOCOS process, an n^+-region is patterned as an active area with silicon nitride. Active area encroachment during the growth of the field oxide can significantly reduce the width of lines. Consequently, patterning an n^+ interconnect requires that the initial layout be larger than the final dimensions to account for bird's beak encroachment. Allowing for this encroachment increases the amount of chip area consumed. The most efficient interconnects are thus the metal and poly layers.

The layout of complicated logic systems is facilitated by having multiple interconnect layers available as needed. Owing to this consideration, process flows have been developed to provide additional conducting layers. Figure 5.35 shows the basic steps used for a double-poly scheme. This gives two polysilicon layers, denoted as Poly I and Poly II. Both may be used as MOSFET gates and interconnect lines. With the final metal layer added, this scheme provides three interconnect levels in addition to n^+ lines.

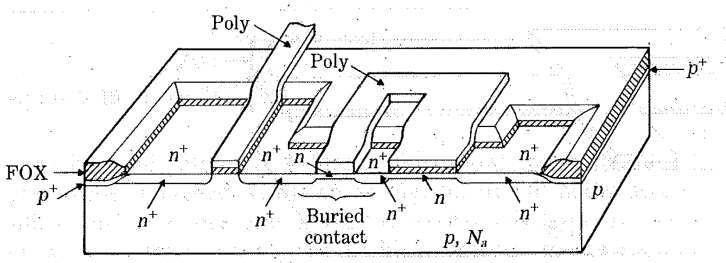
The sequence in Fig. 5.35 demonstrates how two polysilicon layers can be incorporated into the process flow such that either can be used for MOSFET gates. The starting point is the active area shown in Fig. 5.35(a). The (first) gate oxide with thickness x_{ox} is grown, and then the first poly layer (Poly I) is



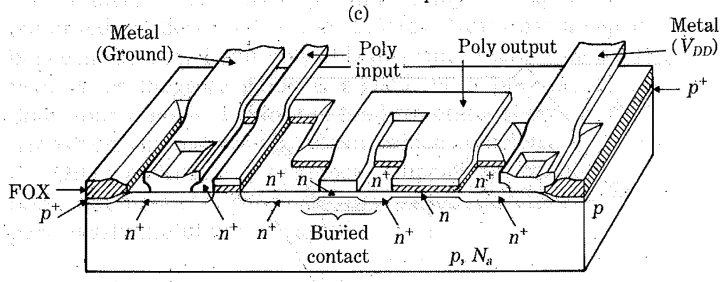
After depletion implant
(a)



Selective gate oxide etch
(b)

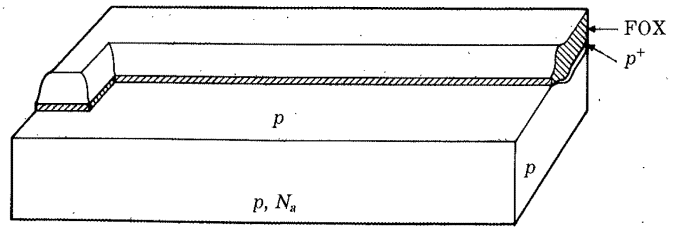


Poly deposition/patterning,
oxide etch and n^+ implant
(c)

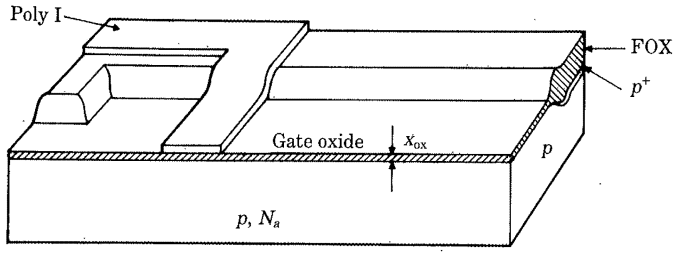


After metallization
(d)

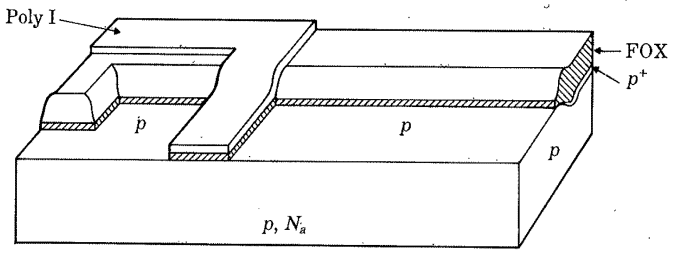
FIGURE 5.34 Formation of a buried poly/n^+ contact.



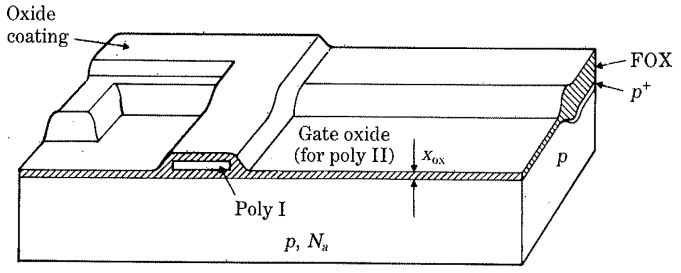
Initial active area
(a)



After gate oxide growth and first-level poly (poly I) deposition/patterning
(b)

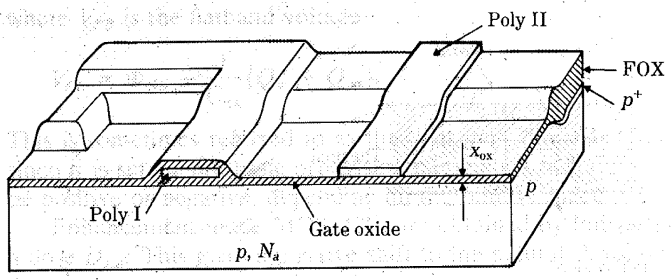


Oxide etch
(c)

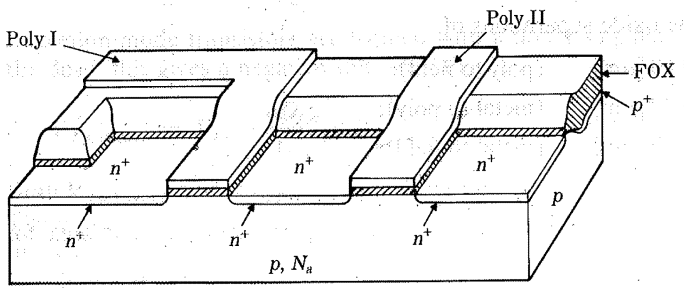


Oxidation for second poly MOSFET gate oxide
(d)

FIGURE 5.35 Double-poly process flow.



Poly II deposition/patterning
(e)



Oxide etch, n^+ implant
(f)

Figure 5.35 (*contd.*)

deposited and patterned. Figure 5.35(b) shows the structure at this point. The gate oxide is etched as drawn in Fig. 5.35(c) to allow access to the p-type silicon surface in preparation for the second poly layer. A second gate oxide is grown to a thickness x_{ox}. In an ideal process, the two gate oxides will be identical. As seen in Fig. 5.35(d), this results in a polyoxide growth over the Poly I layer. The second polysilicon layer, Poly II, is then deposited and patterned as illustrated in Fig. 5.35(e). Note that a capacitor can be formed between the Poly I and Poly II layers since there is an oxide layer between the two. After the gate oxide is etched, both poly layers are exposed. An n^+ drain-source implant gives the final structure, shown in Fig. 5.35(f). The metal deposition and patterning follows, which completes the process flow.

Other multilevel interconnect schemes are common. For example, single-poly, double-metal processes have been developed. Refractory metals are particularly useful in this regard.

EXAMPLE 5.4-1

The important parasitic interconnect capacitances are established during the fabrication cycle. Figure E5.1 shows the layering for a single-poly, single-metal

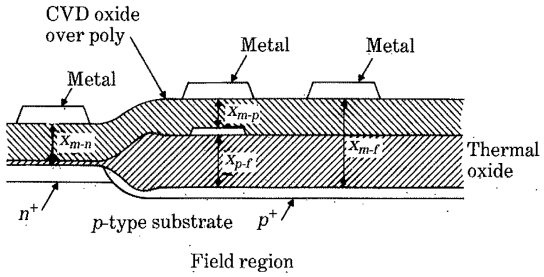


FIGURE E5.1

process. Assume oxide separations of

$$x_{p\text{-}f} = 0.60\,[\mu\text{m}] \quad \text{(poly to field);}$$
$$x_{m\text{-}p} = 0.40\,[\mu\text{m}] \quad \text{(metal to poly);}$$
$$x_{m\text{-}f} = 1.00\,[\mu\text{m}] \quad \text{(metal to field);}$$
$$x_{m\text{-}n^+} = 0.42\,[\mu\text{m}] \quad \text{(metal to } n^+\text{)}$$

as representative values. Compute the oxide capacitances per unit area for each case.

Solution

Since $\varepsilon_{ox} = (3.9)(8.85 \times 10^{-14})$, the interconnect capacitance is calculated using $C_{int} \simeq (3.45 \times 10^{-11})/x_{int}$ [F/cm^2]. This gives

$$C_{p\text{-}f} \simeq 5.76 \times 10^{-9}\,[\text{F/cm}^2] = 0.0576\,[\text{fF}/\mu\text{m}^2],$$
$$C_{m\text{-}p} \simeq 8.63 \times 10^{-9}\,[\text{F/cm}^2] = 0.0863\,[\text{fF}/\mu\text{m}^2],$$
$$C_{m\text{-}f} \simeq 3.45 \times 10^{-9}\,[\text{F/cm}^2] = 0.0345\,[\text{fF}/\mu\text{m}^2],$$
$$C_{m\text{-}n^+} \simeq 8.22 \times 10^{-9}\,[\text{F/cm}^2] = 0.0822\,[\text{fF}/\mu\text{m}^2].$$

These values will be used as typical capacitance levels in later examples. ■

5.5 Threshold Voltage Adjustment

The depletion-load nMOS fabrication sequence described in the last section has associated with it three distinct MOSFET threshold voltages. Multiple-threshold process flows introduce extra degrees of freedom, making them particularly useful in circuit design.

The basic threshold voltage is that obtained from the MOS system physics before ion implantation. From Section 1.1, this is given as

$$V_{T0} = V_{FB} + 2|\phi_F| + \frac{1}{C_{ox}}\sqrt{2q\varepsilon_{Si}N_a(2|\phi_F|)}, \qquad \textbf{(5.5-1)}$$

where V_{FB} is the flatband voltage

$$V_{FB} = \Phi_{GS} - \frac{1}{C_{ox}}(Q_{ss} + Q_{ox}). \tag{5.5-2}$$

This is sometimes referred to as the "natural" threshold voltage of the wafer since it is set by the basic processing parameters including x_{ox} and N_a. It may be positive or negative, depending on the relative magnitude of each term.

Enhancement-mode MOSFETs are obtained by implanting acceptors with a dose $D_{I,a}$. This gives a positive shift to the natural threshold voltage, so

$$V_{T0,\text{E-mode}} = V_{T0} + \frac{qD_{I,a}}{C_{ox}} > 0. \tag{5.5-3}$$

Depletion-mode transistors are formed with a donor implant. Letting $D_{I,d}$ be the dose, this gives a negative shift such that

$$V_{T0,\text{D-mode}} = V_{T0} - \frac{qD_{I,d}}{C_{ox}} < 0. \tag{5.5-4}$$

Both $V_{T0,\text{E-mode}}$ and $V_{T0,\text{D-mode}}$ can be set by adjusting the dose of the implant. These serve as the fundamental circuit design values in the depletion-load process.

Zero-threshold MOSFETs are defined to have threshold voltages slightly greater than 0 [V]. These enhancement-mode devices require only small values of $V_{GS} > 0$ to induce conduction. Zero-threshold transistors are particularly useful for reducing threshold voltage losses and find wide application in certain types of switching and logic control networks.

Zero-threshold transistors can be fabricated by introducing another ion implantation step into the process flow. Alternatively, the natural threshold voltage V_{T0} may itself be slightly positive, which then gives a zero-threshold MOSFET without additional processing. To study this possibility, suppose that the gate is doped n-type with a donor density of $N_{d,\text{poly}}$. The flatband voltage is

$$V_{FB} \simeq -\left(\frac{kT}{q}\right)\ln\left(\frac{N_{d,\text{poly}}}{N_a}\right) - \frac{1}{C_{ox}}(Q_{ss} + Q_{ox}), \tag{5.5-5}$$

where Φ_{GS} from Problem 1.2 in Chapter 1 has been used. Since $V_{FB} < 0$, attainment of a zero-threshold device requires that

$$2|\phi_F| + \frac{1}{C_{ox}}\sqrt{2q\varepsilon_{Si}N_a(2|\phi_F|)} \geq |V_{FB}|. \tag{5.5-6}$$

Assuming that N_a is preset (to give a safe pn^+ reverse breakdown voltage), x_{ox} becomes the critical processing parameter. The value of x_{ox} needed to satisfy eqn. (5.5-6) depends on the oxide charges Q_{ss} and Q_{ox} and also on the doping levels in the gate and substrate. This may or may not be possible.

State-of-the-art processing lines have reached a significant level of complexity. Consequently, additional processing steps that enhance the circuit

capabilities do not present any major technological problems. Economics is the important factor in the final decision.

5.6 nMOS Design Rules and Layout

Chip patterning represents the silicon implementation of a circuit design. Functionality of the chip depends on faithful reproduction of the pattern. *Design rules* are a set of specifications for the mask patterns used in the layout and provide geometric information such as the minimum widths and spacings for each layer. In addition, vertical alignment problems are addressed by specifications on the spacing between lines on different layers. Electrical characteristics are usually provided with the geometric patterning information. Items of interest include k' values, sheet resistances, and capacitance levels.

Design rules provide guidelines for chip layout at the mask level. The specific rules result from a detailed analysis of a given process flow and reflect the limitations of the processing line. In an industrial environment, design rule sets are highly proprietary because an enormous amount of information can be extracted from them. This section deals with the origins and application of design rules in chip mask design and layout.

5.6.1 Physical Basis of Design Rules

Formulation of a design rule set is based on many considerations. The most crucial items are (a) limitations in the lithography, (b) physics of the process flow, and (c) electrical characteristics of the final structures. A detailed discussion of design rule derivation is beyond the scope of this book. However, it is possible to introduce some of the factors that yield the final numbers.

The concept of a *minimum linewidth* is usually associated with limitations in the lithographic resolving power. A minimum linewidth (in units of microns [μm]) specifies the smallest width of a line that can be reliably patterned on a given layer. In terms of the interconnect example shown in Fig. 5.36, this design rule specifies the smallest value of w permitted. Other physical problems are also addressed by this design rule. For example, with a resistivity ρ [Ω-cm] and a layer thickness h, the sheet resistance of the interconnect is

$$R_s = \frac{\rho}{h}[\Omega], \tag{5.6-1}$$

so that

$$R = R_s n\,[\Omega] \tag{5.6-2}$$

gives the resistance of a path with length $\ell = nw$, where n is the number of squares of dimension ($w \times w$) in the direction of current flow. For long interconnects, n may become excessively large. Design rule sets may specify a larger value for min (w) in this case to reduce the overall resistance.

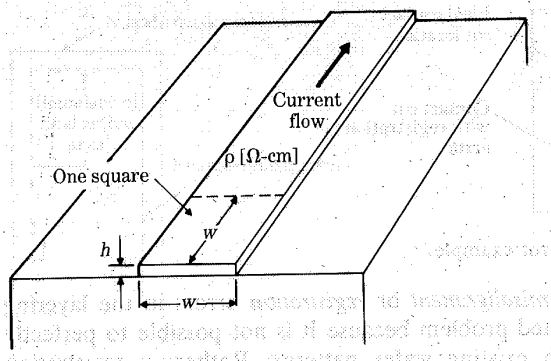


FIGURE 5.36 Interconnect geometry.

Minimum spacing rules for lines on a given layer also tend to originate from the lithography. In Fig. 5.37, two poly lines are separated by a distance d; the design rule for this layer will specify the minimum d permitted. This choice affects the electrical properties of the final structure since a parasitic coupling capacitance C_{coup} exists between the two lines. C_{coup} increases with decreasing separation distance d. The line-to-line spacing thus affects the performance of the circuit. Mutual inductance also enters into the problem since magnetic coupling is present.

Minimum spacing rules between lines on different layers are extremely important in layout. Restrictions arise because the layers must be stacked to form devices. If the vertical alignment of the layers is off, the fabricated structure may not behave properly. The complexity of the problem can be seen by noting that several layers must be applied and patterned to form an integrated circuit. Failure to correctly align even a single layer can result in a totally nonfunctional wafer.

Design rule specifications for spacing between lines of different layers are

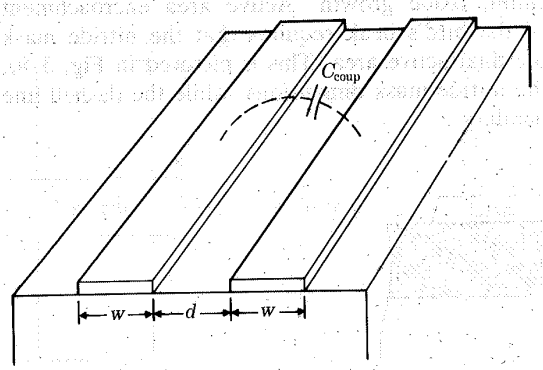


FIGURE 5.37 Interconnect spacing geometry.

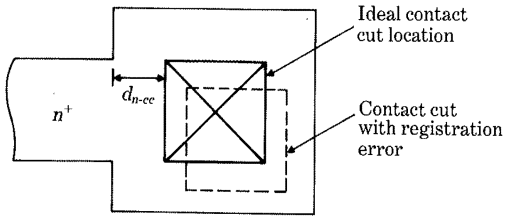


FIGURE 5.38 Registration error example.

chosen to compensate for *misalignment* or *registration* errors in the layering process. This is a complicated problem because it is not possible to perfectly align each new layer with existing wafer patterns. Rather, a registration tolerance must be allowed. This is usually determined by the lithographic equipment and the scheme used to place registration marks on the wafer for subsequent layers.

An important example of the alignment problem is shown in Fig. 5.38. A contact cut is needed in the oxide cover to access the n^+ patterned layer. The ideal location is shown by the solid line. The distance $d_{n\text{-}cc}$ between the edge of the n^+ boundary and the contact cut must be large enough to allow for registration errors when designing the contact cut mask. As long as the actual contact cut is within the n^+ borders, the connection made through the cut will work. The dashed line shows a case where a good cut will be made even though misalignment has occurred.

Spacings between lines on different layers also affect the electrical characteristics. Parasitic coupling capacitance is particularly important. Figure 5.39 shows a poly line and a metal line that are adjacent and thus experience field interactions. The spacing must be large enough to reduce this coupling to a negligible level.

MOSFET layout involves some special considerations. Consider first the active area definition in a LOCOS process. This mask sets the dimensions of the nitride pad used to inhibit oxide growth. Active area encroachment induced by the formation of the bird's beak requires that the nitride mask dimensions be larger than the final active area. This is pictured in Fig. 5.40, where the solid line shows the nitride mask dimensions, while the dashed line gives the final active area boundary.

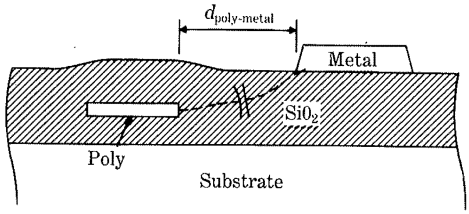


FIGURE 5.39 Poly-metal spacing example.

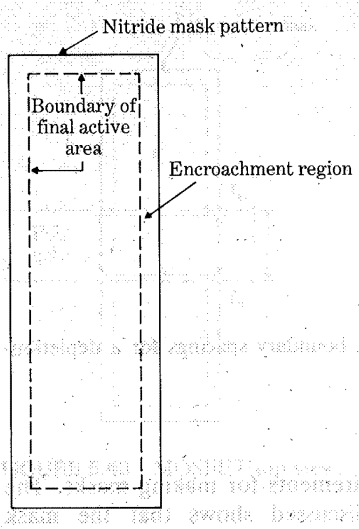


FIGURE 5.40 Nitride mask design.

The poly gate *overhang* distance d_{oh} is shown in Fig. 5.41. The distance d_{oh} is included to ensure that the n^+-regions will be separated in the event of a misalignment between the active area and the poly mask. d_{oh} is crucial in a self-aligned gate process since failure of the poly to cross the active area will result in shorted n^+-regions.

Another example is the mask defining the donor (n-type) implant in a depletion-mode MOSFET. The implanted region must be larger than the device active area to ensure that an n-layer is created if misalignment occurs. The outline of the depletion implant is shown in Fig. 5.42 by the dashed line. Similar problems arise for selectively implanted E-mode transistors.

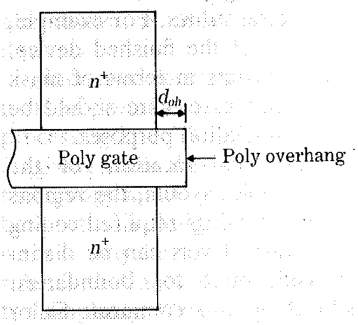


FIGURE 5.41 Poly gate overhang in a MOSFET.

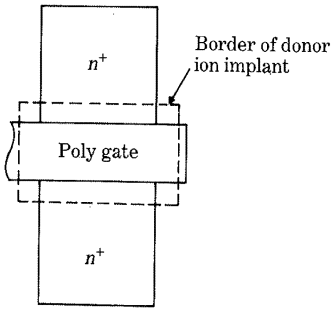


FIGURE 5.42 Threshold adjustment ion implant boundary spacings for a depletion-mode MOSFET layout.

5.6.2 Mask Drawings

Design rules are spacing and linewidth requirements for making masks. The active area encroachment problem just discussed shows that the mask geometries are different from the final device layout. The mask pattern set is sometimes referred to as the *drawn* layout, while the resulting chip patterns are termed the *actual* or *final* layouts. Encroachment, lateral diffusion, isotropic etching, and other physical processes require that a distinction be made.

An example of this was presented in the discussion of the overlap capacitance C_o in Section 4.6. Overlapping exists because of lateral doping effects in a self-aligned MOSFET structure. Denoting the drawn channel length by L', this gave

$$L' = L + 2L_o, \tag{5.6-3}$$

where L is the actual channel length. Figure 5.43 shows the superposed mask patterns for the gate and active area in solid lines. The final n^+ geometry (after processing) is indicated by the dashed lines. The boundary around the device edges is affected by both encroachment and lateral doping effects.

Device characteristics depend on the final geometric values. For example, L and W in I_D always refer to the length and width of the finished device. However, it is usually more convenient to present layouts in terms of mask drawings, particularly during the design phase. In this case, care should be exercised when obtaining geometric data for circuit simulation purposes.

Mask-level layout drawings usually show the ideal locations of the patterns. Multiple-layer drawings are common. In simple layouts, the regions are labeled (e.g., poly, n^+, etc.). More complicated drawings required coding to keep the layers separated. In single-color drawings, levels can be distinguished using different line types (solid, dashed, etc.) for boundaries. Alternatively, variations in shading or cross-hatching are common. Color coding is useful for CAD displays and plotters and is particularly nice for *stick*

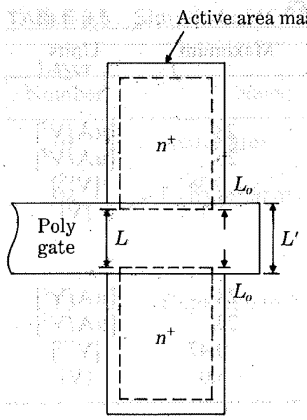


FIGURE 5.43 MOSFET top view.

diagrams in VLSI system design. (A stick diagram displays only the routing of the patterns.)

5.6.3 Design Rule Sets

Design rules are process-specific, so minimum linewidths, spacings, dimensions, etc., are given in units of microns [μm]. The actual numbers provide a measure of the integration density since small values indicate higher packing densities. A design rule set can easily be 50 pages or longer. Simplified sets exist for applications that do not maximize packing density. An example of a basic nMOS design rule set is provided in Table 5.4. This is a generic list compiled for illustration only. Levels (masking layers) are denoted by number (01, 02, etc.) and name ("Active area," "Enhancement implant," etc.), with critical dimensions provided for each. A typical list of electrical parameters is also shown. Note in particular the variations in k' values for different-sized devices.

Increased interest in IC chip design for VLSI has prompted the development of *portable* design rules that can, in principle, be adapted to an arbitrary processing line. These are formulated in terms of a fundamental design length λ (giving them the name λ *design rules*). The length λ is chosen according to technology and is a measure of the minimum resolution of the lithography. A base set of λ design rules is provided in Table 5.5. Portability is incorporated since λ may be varied as needed. The main disadvantage of portable design rules is that most spacings are chosen as integer or half-integer multiples of λ for simplicity. This generally precludes the possibility of attaining a maximum packing density.

TABLE 5.4 Simplified nMOS Design Rule Set Example

Electrical Characteristics

Parameter	Minimum	Typical	Maximum	Units
1. Enhancement mode				
k' ($L > 5$ [μm])	22	25	28	[$\mu A/V^2$]
k' ($L < 5$ [μm])	20	22	26	[$\mu A/V^2$]
γ	0.31	0.35	0.38	[$V^{1/2}$]
V_{T0}	0.7	0.9	1.1	[V]
2. Depletion mode				
k' ($L > 5$ [μm])	22	25	28	[$\mu A/V^2$]
k' ($L < 5$ [μm])	22	25	28	[$\mu A/V^2$]
γ	0.37	0.43	0.47	[$V^{1/2}$]
V_{T0}	−3.7	−3.4	−3.0	[V]

Layout Rules

Layer Number	Name	Rules (Units of [μm])	
01	Active area	Minimum dimension	3 × 4
		Minimum spacing	3
02	Enhancement implant	Minimum overlap with active area (FOX)	3
		Minimum overlap with active area (n^+)	2
03	Depletion implant	Minimum overlap with active area (FOX)	3
		Minimum overlap with active area (n^+)	3
		Minimum spacing from E-mode MOSFET	3
04	Buried contact	Minimum overlap	4
		Minimum spacing to poly or device	3
05	Poly	Mininum width	3
		Minimum spacing	3
		Minimum MOSFET gate overhang	2
		Minimum spacing to n^+	3
		Minimum dimensions for buried contact	3 × 4
06	Metal contact mask	Minimum dimensions	3 × 3
		Minimum spacing to active area edge	3
		Minimum spacing to poly edge	3
07	Metal	Minimum width	6
		Minimum spacing	4
08	Passivation (overglass)		

TABLE 5.5 Simplied nMOS Portable Layout Rule Set Example

Layer Number	Name	Rules	
01	Active area	Minimum dimension	$2\lambda \times 2\lambda$
		Minimum spacing	2λ
02	Enhancement implant	Minimum overlap with active area (FOX)	1.5λ
		Minimum overlap with active area (n^+)	1.5λ
03	Depletion implant	Minimum overlap with active area (FOX)	1.5λ
		Minimum overlap with active area (n^+)	1.5λ
		Minimum spacing from E-mode MOSFET	2λ
04	Buried contact	Minimum overlap	2λ
		Minimum spacing to poly or device	2λ
05	Poly	Minimum width	2λ
		Minimum spacing	2λ
		Minimum MOSFET gate overhang	2λ
		Minimum spacing to n^+	1λ
		Minimum dimensions for buried contact	$2\lambda \times 4\lambda$
06	Metal contact mask	Minimum dimensions	$2\lambda \times 2\lambda$
		Minimum spacing to active area edge	2λ
		Minimum spacing to poly edge	2λ
07	Metal	Minimum width	3λ
		Minimum spacing	3λ
08	Passivation (overglass)		

5.6.4 Some Comments on Layout

Integrated circuit layout is usually classified as an art that is best learned by practice. If integration density is not important, layout becomes relatively staightforward. The complicating factor is that the geometry determines the circuit parameters such as parasitic capacitance and resistance; these in turn affect the performance and the overall design. Iterations are often required.

Computer-aided design (CAD) and computer-aided engineering (CAE) are mandatory for complete layout control. Varieties of workstations are

available for graphics layout and schematic capture (where the layout can be used to generate a component list). Most layout software tools provide a design rule checker (DRC), which can be used to determine if any design rules have been violated. In addition, circuit compaction and interconnect routing capabilities are common.

The advent of VLSI has spurred the development of CAD/CAE techniques. Advances in these areas have rapidly changed the level of layout complexity possible. Circuit design is intimately related to layout, so maintaining pace with the state of the art is mandatory.

5.7 Processing Variations

Variations in the process flow result in nonuniformities on the finished wafers. Electrical and structural differences can be detected when comparing wafers in a given lot. For that matter, die characteristics vary across a single wafer. Precise control of the fabrication process is not possible.

Geometric variations are addressed by adhering to a design rule set in the layout. Electrical variations directly affect the circuit performance, so a circuit designer must work around the processing variables to ensure working circuits. This complicates matters since the design methodology must be extended and refined to a level that is compatible with the process limitations.

5.7.1 Generalized Modeling

Process variations are usually analyzed using statistical techniques. The problem can be understood by referring to the example in Fig. 5.44, which shows a histogram for threshold voltages in a given test group. The "target"

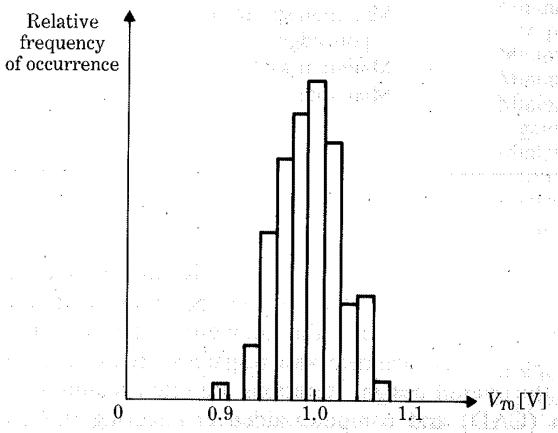


FIGURE 5.44 Histogram example for threshold voltage.

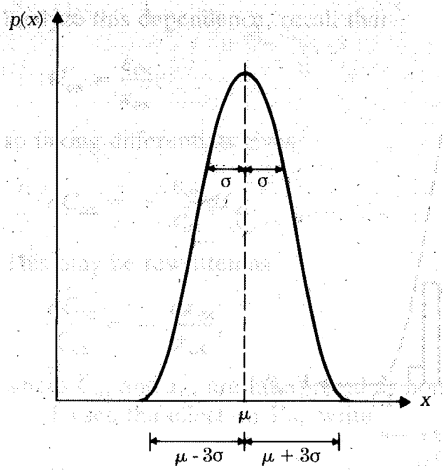


FIGURE 5.45 Gaussian probability distribution density.

value of the process is assumed to be $V_{T0} = +1$ [V]; this is termed the *nominal* or *working* value. Design rules generally provide nominal values for electrical characteristics.

Modeling the variations of a parameter is complicated and can be approached in several different ways. The simplest analysis is based on Gaussian distributions. For a given parameter x, the probability density $p(x)$ is given by

$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-[(x-\mu)^2/2\sigma^2]}, \tag{5.7-1}$$

such that $p(x)\,dx$ gives the probability of finding the parameter value in the range x and $(x + dx)$. In this equation, μ is the mean or average value of the parameter and σ is the *standard deviation* of the distribution; σ measures the *spread* of the distribution. Figure 5.45 shows the general properties of a Gaussian.

Probability densities are used to compute probabilities. To calculate the probability P of finding the variable x between X_1 and X_2, differential contributions $p(x)\,dx$ in this range are summed by integrating:

$$P = \int_{X_1}^{X_2} p(x)\,dx. \tag{5.7-2}$$

Since the total probability of measuring x in $[-\infty, +\infty]$ is unity, the distribution is *normalized* to 1 by requiring

$$\int_{-\infty}^{+\infty} p(x)\,dx = 1. \tag{5.7-3}$$

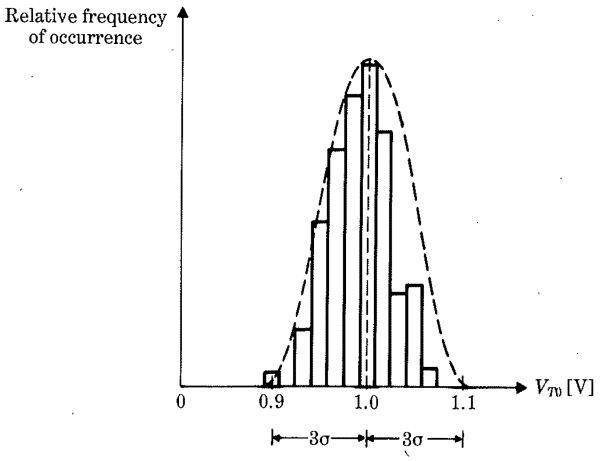


FIGURE 5.46 Adaptation of Gaussian distribution to threshold voltage histogram with large variations.

These equations indicate that the probability of measuring a value in a given range is proportional to the area under the curve.

It is useful to reference the probability to the values of μ and σ. An analysis of the problem shows that

$$\int_{\mu-3\sigma}^{\mu+3\sigma} p(x)\,dx \simeq 0.997, \tag{5.7-4}$$

which indicates that about 99.7% of the sample values lie in the interval $(\mu - 3\sigma, \mu + 3\sigma)$. The quantity 6σ is a convenient measure of the total spread.

To apply the Gaussian distribution to the threshold voltage problem in Fig. 5.44, let the nominal value $V_{T0} = 1$ [V] represent the mean. The standard deviation σ can be obtained by fitting the histogram to a Gaussian, as shown in Fig. 5.46. This allows for an estimation of $3\sigma \equiv \Delta V_{T0}$, so the threshold voltage can be assumed to lie in the range

$$V_{T0} = 1 \pm (\Delta V_{T0})\,[\text{V}]. \tag{5.7-5}$$

This becomes a circuit design constraint.

5.7.2 Analysis of Device Parameters

Process variations directly affect the device parameters used in circuit design. As an example, consider the threshold voltage spread just described. Variations in the gate oxide thickness x_{ox} are an important factor in this problem. To

analyze this dependence, recall that

$$C_{ox} = \frac{\varepsilon_{ox}}{x_{ox}}, \tag{5.7-6}$$

so taking differentials gives

$$dC_{ox} = -\frac{\varepsilon_{ox}}{x_{ox}^2} dx_{ox}. \tag{5.7-7}$$

This may be rewritten as

$$\frac{dC_{ox}}{C_{ox}} = -\frac{dx_{ox}}{x_{ox}}, \tag{5.7-8}$$

where C_{ox} and x_{ox} are interpreted as nominal values.

To see the effect on V_{T0}, write

$$V_{T0} = V_{FB} + 2|\phi_F| + \frac{x_{ox}}{\varepsilon_{ox}}|Q_{B0}| + \frac{qD_I x_{ox}}{\varepsilon_{ox}}. \tag{5.7-9}$$

Assuming for the moment that only variations in x_{ox} are important, taking differentials yields

$$dV_{T0} = \frac{\partial V_{FB}}{\partial x_{ox}} dx_{ox} + \frac{1}{\varepsilon_{ox}}|Q_{B0}|\, dx_{ox} + \frac{qD_I}{\varepsilon_{ox}} dx_{ox}. \tag{5.7-10}$$

The flatband voltage term evaluates to (see eqn. 1.1-22)

$$\frac{\partial V_{FB}}{\partial x_{ox}} dx_{ox} = -\frac{1}{\varepsilon_{ox}}(Q_{ss} + Q_{ox})\, dx_{ox}, \tag{5.7-11}$$

so dV_{T0} can then be written in the form

$$dV_{T0} = [Q_{ss} + Q_{ox} - qD_I - |Q_{B0}|]\frac{dC_{ox}}{C_{ox}^2}, \tag{5.7-12}$$

showing the dependence on variations in C_{ox}. This is useful because C_{ox} can be measured in the lab.

The general analysis technique can be extended to include more than one variable. For example, to compute the variation in V_{T0} due to changes in both x_{ox} and D_I, write

$$dV_{T0} = \frac{\partial V_{T0}}{\partial x_{ox}} dx_{ox} + \frac{\partial V_{T0}}{\partial D_I} dD_I. \tag{5.7-13}$$

This adds a term

$$+\left(\frac{qD_I}{C_{ox}}\right)\frac{dD_I}{D_I} \tag{5.7-14}$$

to eqn. (5.7-13). For an arbitrary function $f = f(A, B, C, \ldots)$, variations due

to the parameters $A, B, C, \ldots$, can be computed by

$$df = \left(\frac{\partial f}{\partial A}\right) dA + \left(\frac{\partial f}{\partial B}\right) dB + \left(\frac{\partial f}{\partial C}\right) dC + \cdots. \tag{5.7-15}$$

To apply these results to a realistic situation, approximate

$$df \simeq \Delta f, \qquad dA \simeq \Delta A, \qquad dB \simeq \Delta B, \qquad \ldots, \tag{5.7-16}$$

where (Δf) represents the spread in f values due to changes (ΔA), (ΔB), etc., in the processing line.

5.7.3 Circuit Design Criteria

Process variations enter into the circuit design problem because they affect the voltages and currents. Consider a depletion-load nMOS inverter. The driver-load ratio for a given V_{OL} is calculated from

$$\beta_R = \frac{k'_D(W/L)_D}{k'_L(W/L)_L} = \frac{|V_{TL}(V_{OL})|^2}{[2(V_{OH} - V_{TD})V_{OL} - V_{OL}^2]}, \tag{5.7-17}$$

where

$$V_{TL}(V_{OL}) = V_{TOL} + \gamma_L(\sqrt{V_{OL} + 2|\phi_F|} - \sqrt{2|\phi_F|}). \tag{5.7-18}$$

Normal processing variations exist in the threshold voltages V_{TD} and V_{TL} and also in the k' values. To complicate matters more, power supply variations (as manifest in V_{OH}) should also be accounted for in the circuit design. A good design will account for these possibilities such that the range of V_{OL} output voltages is acceptable when interfaced with the other circuitry.

The simplest approach to designing around process variations is to use nominal values in the initial design cut and then vary the important parameters to track the circuit performance. Although a few hand calculations can be made, this generally requires a computer simulation of the circuit. For example, SPICE provides for a sensitivity analysis (.SENS) to determine tolerances. Experience is probably the best teacher for this type of problem.

Chip yield is also related to the problem. A processed wafer contains many die, only a fraction of which will operate correctly. The yield Y is defined to be the percentage of good die on the wafer. Defective chips either will be completely nonfunctional or may not meet electrical specifications for the product. In a new design, this may result from problems in processing, circuit design, or both. The importance of this consideration is immediately obvious: low yields give low (or negative) profits! Economics thus dictates that a circuit designer (who wants to stay employed) take variations into account.

5.8 CMOS Technologies

CMOS fabrication is complicated by the fact that both n-channel and p-channel MOSFETs are used. This requires that bulk regions of both

polarities be available at the wafer surface. nMOS transistors are fabricated in grounded p-type bulk regions, while pMOS transistors are made in n-type bulk regions that are electrically connected to the power supply. The substrate wafer can be chosen either n-type or p-type. The CMOS process flow must include a step to create regions of the opposite polarity.

A possible approach to CMOS is to use a p-type wafer and selectively dope *n-tub* regions for placement of p-channel MOSFETs. The primary sequences for an inverter created in an n-tub LOCOS flow are shown in Fig. 5.47. The starting point is a p-type wafer. Ion implantation or diffusion is used to create n-tub regions; this is illustrated in Fig. 5.47(a). Allowance has been made for eventual connection of the n-tub region (the bulk electrode for pMOS transistors) to the power supply V_{DD} rail. The next step defines the active areas by nitride patterning, and then the LOCOS FOX is grown. Figure 5.47(b) shows the structure at this point. Note that the different widths for the two active areas have been patterned to give a symmetric inverter with $\beta_n = \beta_p$.

Gate oxide regrowth comes next, and then poly is deposited for the gate layer. Poly patterning followed by selective n^+ and p^+ implants gives the chip surface shown in Fig. 5.47(c). After a CVD oxide is applied and patterned, metal is deposited to give the output, V_{DD}, and ground connections shown in Fig. 5.47(d). This completes the main processing steps in the LOCOS flow.

Other approaches are possible. The direct complement of the n-tub process is the *p-tub* flow where an n-type wafer is used for pMOS transistors and p-tub regions are provided for n-channel MOSFETs. p-tub technology is ideal for retrofitting a pMOS (only) line for CMOS. *Twin-tub* processes have also been developed. In this approach, a thin CVD *epitaxial* layer of lightly doped silicon is grown on the substrate wafer. Then individual n-regions and p-regions are patterned in the epitaxial layer to provide MOSFET active areas.

A significant problem that can occur in CMOS circuits is *latch-up*. Figure 5.48(a) shows the origin of the problem in the n-tub inverter. Tracking the silicon layers from the power supply V_{DD} to ground shows the existence of a *4-layer* p^+npn^+ structuring. This can be interpreted in different ways. Those familiar with power semiconductor devices will recognize this as the basis for a *silicon-controlled rectifier* (SCR). An SCR acts like a diode but is triggered into conduction using an external electrode. Alternately, the layering can be visualized as a pair of opposite polarity bipolar transistors. This view point is adopted here and gives the *npn/pnp* combination shown in Fig. 5.48(b).

Latch-up describes the condition in which current flow is diverted from the power supply to the ground electrode through the p^+npn^+ layering. Once current flow is initiated, the only way to stop it is to disconnect the power supply. Since no current is supplied to the circuit, the logic will not operate. In the worst-case scenario, the chip burns up from excessive Joule heating.

The bipolar transistor equivalent circuit shows the substrate current I from V_{DD}. If the base-emitter junction of the *pnp* transistor becomes forward biased, I begins to flow. This provides base current I_{Bn} to the *npn* transistor, which turns the device on. Collector current $I_{Cn} = I_{Bp}$ from the *npn* transistor

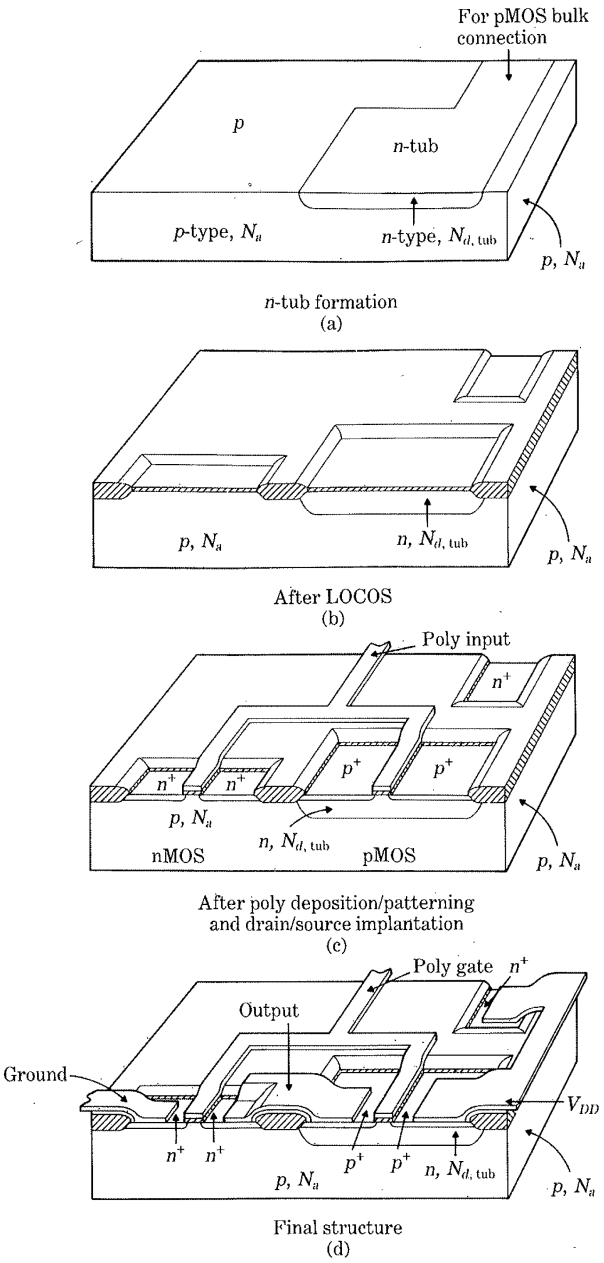


FIGURE 5.47 CMOS process flow.

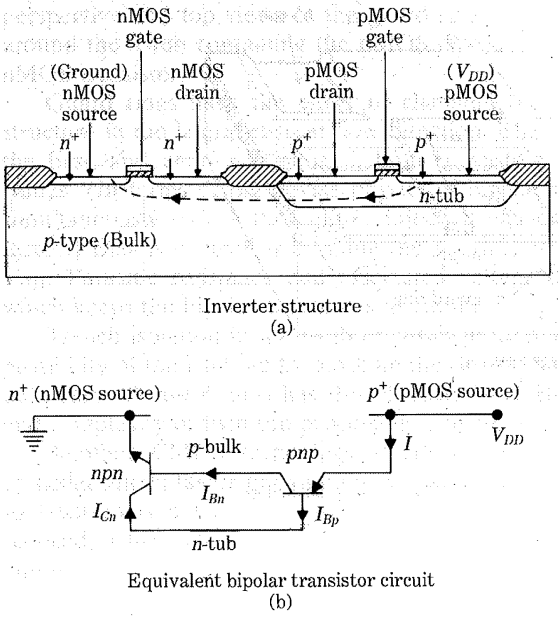


FIGURE 5.48 Latch-up in an *n*-tub CMOS inverter.

forces the *pnp* transistor to conduct more current. Feedback thus sustains the current flow and sinks current directly from V_{DD} to ground. This type of action leads to latch-up.

The bipolar analogy can be analyzed for the condition of latch-up. Denoting the common-base current gains by α_{npn} and α_{pnp}, the analysis gives [16]

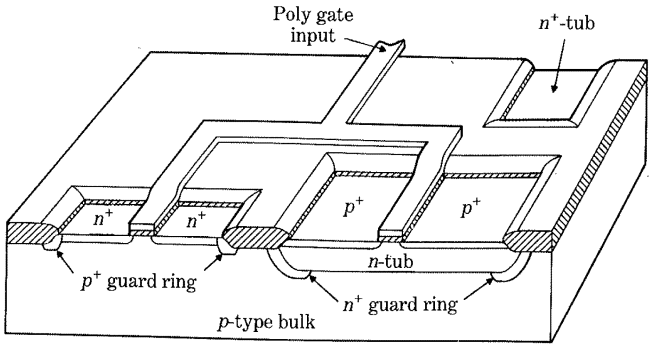
$$I = \frac{I_{C0}}{1 - \alpha_{npn} - \alpha_{pnp}}, \tag{5.8-1}$$

where I_{C0} is the reverse leakage current. The current gains are dependent on current such that $(d\alpha/dI) > 0$. Both α_{npn} and α_{pnp} increase with increasing current, so the condition

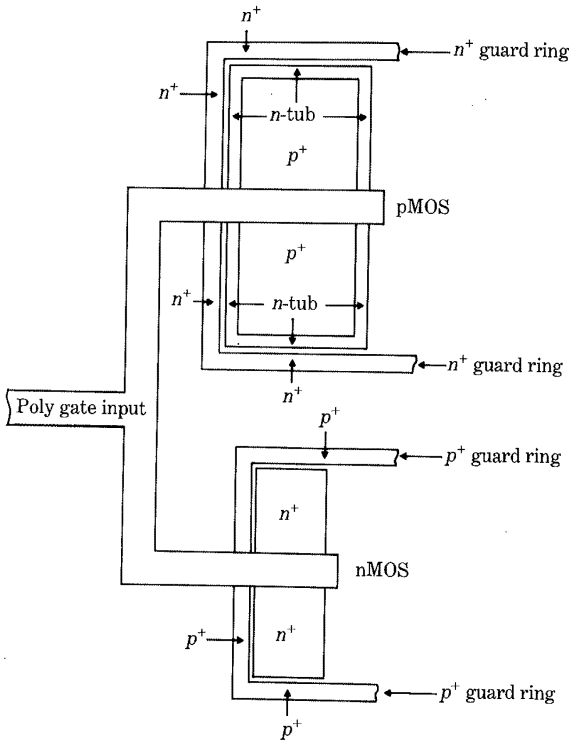
$$\alpha_{npn} + \alpha_{pnp} = 1 \tag{5.8-2}$$

gives a singularity corresponding to latch-up current flow. After this is initiated, feedback sustains current flow.

Different approaches are available for dealing with latch-up. Using the bipolar transistor analogy shows that reducing the common-base current gains of the parasitic transistors can prevent the structure from ever reaching the singularity condition (eqn. 5.8-2). A direct way to implement this is to place heavily doped *guard rings* around the MOSFETs. Figure 5.49 shows both



Chip structuring
(a)



Layout
(b)

FIGURE 5.49 Guard rings for latch-up prevention.

perspective and top views of the guard ring placement. An n^+ ring is placed around the n-tub containing the pMOS device, while a p^+ ring surrounds the nMOS transistor.

Guard rings have the effect of changing the layering into a $p^+n^+p^+n^+$ structure in the lateral current flow direction. This reduces the effectiveness of the base and emitter regions in both transistors, which results in lower α values. (In terms of the device physics of bipolar transistors, the guard rings simultaneously reduce the emitter injection efficiency and the base transport factor.) Bias is controlled by tying the p^+ rings to 0 [V] and the n^+ rings to V_{DD}. Parasitic resistance and associated voltage drops are then eliminated, which keeps the bipolar transistors in cutoff.

Trench isolation in an n-tub or p-tub process automatically precludes the possibility of the latch-up by blocking the current flow path. No guard rings are necessary. Twin-tub also has this characteristic. However, the increased cost and complexity of twin-tub processing is significant.

Another CMOS technology is *silicon on sapphire* (SOS), which uses epitaxial silicon layers grown on a sapphire substrate. Sapphire is an insulator, so latch-up is not a problem. Moreover, the capacitance levels are greatly reduced, which allows for higher-speed circuits. Cost, however, is again a major factor against using SOS for everyday applications.

EXAMPLE 5.8-1

A CMOS process uses n-tub technology with the following specifications:

$$x_{ox} = 500\ [\text{Å}],\quad \mu_n \simeq 580\ [\text{cm}^2/\text{V-sec}],\quad \mu_p \simeq 218\ [\text{cm}^2/\text{V-sec}],$$
$$N_a = 10^{15}\ [\text{cm}^{-3}],\quad N_{a,sw} = 5 \times 10^{15}\ [\text{cm}^{-3}],$$
$$N_{d,\text{tub}} = 10^{16}\ [\text{cm}^{-3}],\quad N_{d,sw} = 6 \times 10^{16}\ [\text{cm}^{-3}],$$
$$N_{d,n^+} = 10^{20}\ [\text{cm}^{-3}],\quad N_{a,p^+} = 10^{20}\ [\text{cm}^{-3}]\quad (n^+ \text{ and } p^+ \text{ regions}).$$

The base threshold voltages are set at $V_{T0n} = +0.70\ [\text{V}]$, $V_{T0p} = -0.80\ [\text{V}]$. Calculate the important circuit design parameters.

Solution

First,

$$C_{ox} = \frac{(3.9)(8.854 \times 10^{-14})}{0.05 \times 10^{-4}} \simeq 6.91 \times 10^{-8}\ [\text{F/cm}^2].$$

Next,

$$k_n' \simeq (580)(6.91 \times 10^{-8}),\qquad k_n' \simeq 40\ [\mu\text{A/V}^2],$$
$$k_p' \simeq (218)(6.91 \times 10^{-8}),\qquad k_p' \simeq 15\ [\mu\text{A/V}^2].$$

The body bias coefficients are

$$\gamma_n \simeq \frac{\sqrt{2(1.6 \times 10^{-19})(11.8)(8.854 \times 10^{-14})(10^{15})}}{6.91 \times 10^{-8}} \simeq 0.26\,[\mathrm{V}^{1/2}] \quad \text{(nMOS)},$$

$$\gamma_p \simeq \frac{\sqrt{2(1.6 \times 10^{-19})(11.8)(8.854 \times 10^{-14})(10^{16})}}{6.91 \times 10^{-8}} \simeq 0.84\,[\mathrm{V}^{1/2}] \quad \text{(pMOS)}.$$

For an n-channel FET, the bulk Fermi potential is

$$2\,|\phi_{Fp}| \simeq 2(0.026) \ln\left(\frac{10^{15}}{1.45 \times 10^{10}}\right) \simeq 0.579\,[\mathrm{V}] \quad \text{(nMOSFETs)},$$

while a p-channel FET uses

$$2\,|\phi_{Fn}| \simeq 2(0.026) \ln\left(\frac{10^{16}}{1.45 \times 10^{10}}\right) \simeq 0.699\,[\mathrm{V}] \quad \text{(pMOSFETs)}.$$

Next, examine the junction capacitances for the nMOS n^+/p-substrate and pMOS p^+/n-tub drain and source regions.

$$nMOS: \qquad \phi_o \simeq (0.026) \ln\left[\frac{10^{15}10^{20}}{(1.45 \times 10^{10})^2}\right] \simeq 0.879\,[\mathrm{V}],$$

and

$$C_{j0} \simeq \sqrt{\frac{(1.6 \times 10^{-19})(11.8)(8.85 \times 10^{-14})}{2(10^{-15} + 10^{-20})(0.879)}} \simeq 9.75 \times 10^{-9}\,[\mathrm{F/cm^2}],$$

while the sidewall contribution quantities are

$$\phi_{osw} \simeq (0.026) \ln\left[\frac{(5 \times 10^{15})10^{20}}{(1.45 \times 10^{10})^2}\right] \simeq 0.921\,[\mathrm{V}],$$

$$C_{j0sw} \simeq \sqrt{\frac{(1.6 \times 10^{-19})(11.8)(8.854 \times 10^{-14})}{2(2 \times 10^{-16} + 10^{-20})(0.921)}} \simeq 2.13 \times 10^{-8}\,[\mathrm{F/cm^2}].$$

The sidewall capacitance must be multiplied by the junction depth x_{jn} for use in circuit calculations.

$$pMOS: \qquad \phi_o \simeq (0.026) \ln\left[\frac{10^{16}10^{20}}{(1.45 \times 10^{10})^2}\right] \simeq 0.939\,[\mathrm{V}],$$

and

$$C_{j0} \simeq \sqrt{\frac{(1.6 \times 10^{-19})(11.8)(8.854 \times 10^{-14})}{2(10^{-16} + 10^{-20})(0.939)}} \simeq 2.98 \times 10^{-8}\,[\mathrm{F/cm^2}].$$

The sidewall contributions are

$$\phi_{osw} \simeq (0.026) \ln\left[\frac{(6 \times 10^{16})10^{20}}{(1.45 \times 10^{10})^2}\right] \simeq 0.985\,[\mathrm{V}],$$

$$C_{j0sw} \simeq \sqrt{\frac{(1.6 \times 10^{-19})(11.8)(8.854 \times 10^{-14})}{2(1.67 \times 10^{-17} + 10^{-20})(0.985)}} \simeq 7.23 \times 10^{-8}\,[\mathrm{F/cm^2}]. \quad \blacksquare$$

TABLE 5.6 Simplified CMOS Design Rule Set Example (*n*-tub)

Electrical Characteristics

Parameter	Minimum	Typical	Maximum	Units
1. *n*-channel MOSFETs				
k' ($L > 5$ [μm])	36	40	44	[μA/V^2]
k' ($L < 5$ [μm])	34	37	43	[μA/V^2]
γ	0.23	0.26	0.28	[V$^{1/2}$]
V_{T0}	0.7	0.9	1.1	[V]
2. *p*-channel MOSFET				
k' ($L > 5$ [μm])	15	18	21	[μA/V^2]
k' ($L < 5$ [μm])	14	16	20	[μA/V^2]
γ	0.76	0.84	0.93	[V$^{1/2}$]
V_{T0}	−1.0	−0.85	−0.7	[V]

Layout Rules

Layer Number	Name	Rules (Units of [μm])	
01	*n*-tub	Minimum dimensions	3 × 3
		Minimum spacing	3
02	Active area	Minimum dimension	3 × 4
		Minimum spacing	3
		Minimum distance to *n*-tub edge	9
03	nMOS threshold implant	Minimum overlap with active area (FOX)	3
		Minimum overlap with active area (n^+)	2
04	Poly	Minimum width	3
		Minimum spacing	3
		Minimum MOSFET gate overhang	2
		Minimum spacing to n^+	3
05	Metal contact mask	Minimum dimensions	3 × 3
		Minimum spacing to active area edge	3
		Minimum spacing to poly edge	3
06	Metal	Minimum width	6
		Minimum spacing	4
07	Passivation (overglass)		

5.9 CMOS Design Rules

Design rules for CMOS circuits are based on the same considerations as in nMOS designs. Lithography, processing limitations, and final electrical characteristics must be considered. The main differences in formulating a set of CMOS design rules are that (a) an opposite polarity tub mask must be included and (b) guard rings or some latch-up prevention becomes important. Both tend to increase the real estate requirement over that needed for an equivalent nMOS technology (i.e., for the same linewidth). However, as mentioned before, CMOS technologies have been developed to an advanced stage so that a direct comparison with NMOS is no longer meaningful.

CMOS design rules are process-specific. An illustrative set of rules is provided in Table 5.6. Portable design rules can also be formulated. A λ set example is listed in Table 5.7. An example of a CMOS inverter layout is shown in Fig. 5.50.

TABLE 5.7 Simplified CMOS Portable Layout Rule Set Example (*n*-tub)

Layer Number	Name	Rules	
01	*n*-tub	Minimum dimensions	$2\lambda \times 2\lambda$
		Minimum spacing	3λ
02	Active area	Minimum dimension	$3\lambda \times 2\lambda$
		Minimum spacing	2λ
		Minimum distance to *n*-tub edge	5λ
03	nMOS threshold implant		
		Minimum overlap with active area (FOX)	2λ
		Minimum overlap with active area (n^+)	1.5λ
04	Poly	Minimum width	2λ
		Minimum spacing	2λ
		Minimum MOSFET gate overhang	2λ
		Minimum spacing to n^+	2λ
05	Metal contact mask		
		Minimum dimensions	$3\lambda \times 2\lambda$
		Minimum spacing to active area edge	2λ
		Minimum spacing to poly edge	2λ
06	Metal	Minimum width	3λ
		Minimum spacing	3λ
07	Passivation (overglass)		

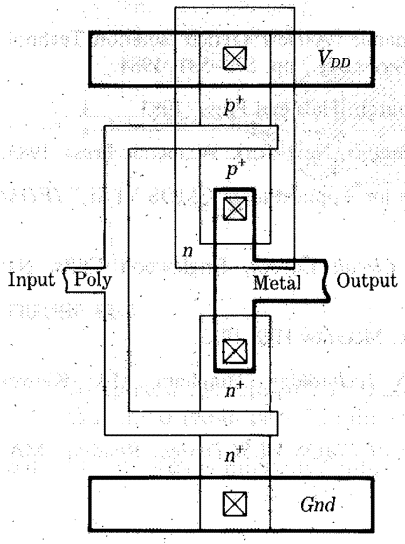


FIGURE 5.50 CMOS inverter layout in an n-tub technology (equal device size).

REFERENCES

(1) P. Antognetti, et al. (eds.), *Process and Device Simulation for MOS-VLSI Circuits* (NATO ASI Series), The Hague: Martinus Nijhoff, 1983.

(2) L. Chang, "A Derivative Method to Determine a MOSFET's Effective Channel Length and Width Electrically," *IEEE Electron Device Lett.*, vol. EDL-7, pp. 229–231, 1986.

(3) R. A. Colclaser, *Microelectronics Processing and Device Design,* New York: Wiley, 1980.

(4) L. Esaki and G. Soncini (eds.), *Large Scale Integrated Circuits Technology* (NATO ASI Series), The Hague: Martinus Nijhoff, 1982.

(5) S. K. Ghandhi, *VLSI Fabrication Principles,* New York: Wiley, 1983.

(6) A. B. Glaser and G. E. Subak-Sharpe, *Integrated Circuit Engineering,* Reading, MA: Addison-Wesley, 1977.

(7) L. A. Glasser and D. W. Dobberpuhl, *The Design and Analysis of VLSI Circuits,* Reading, MA: Addison-Wesley, 1985.

(8) D. J. Hamilton and W. G. Howard, *Basic Integrated Circuit Engineering,* New York: McGraw-Hill, 1975.

(9) G. De Mey, "Stochastic Geometry Effects in MOS Transistors," *IEEE J. Solid-State Circuits,* vol. SC-20, pp. 865–870, 1985.

(10) H. Mikoshiba, T. Homma, and K. Hamano, "A New Trench Isolation Technology as a Replacement for LOCOS," *IEDM Tech. Dig.*, pp. 578–581, 1984.

(11) A. D. Milne (ed.), *MOS Devices,* Edinburgh: Halstead Press, 1983.

(12) S. P. Murarka, *Silicides for VLSI Applications,* New York: Academic Press, 1983.

(13) R. D. Rung, "Trench Isolation Prospects for Application in CMOS VLSI," *IEDM Tech. Dig.*, pp. 574–577, 1984.

(14) B. Spinks, *Introduction to Integrated Circuit Layout,* Englewood Cliffs, NJ: Prentice-Hall, 1985.

(15) S. M. Sze, *VLSI Technology,* New York: McGraw-Hill, 1983.

(16) R. R. Troutman, *Latchup in CMOS Technology,* Hingham, MA: Kluwer Academic Publishers, 1986.

(17) N. Weste and K. Eshraghian, *Principles of CMOS VLSI Design,* Reading, MA: Addison-Wesley, 1985.

PROBLEMS

5.1 A thermal oxide layer is grown on a (100) silicon wafer using two steps. The first is a steam oxidation at 1000 [°C] for 60 minutes. This is followed by a dry O_2 growth at 1000 [°C] for 45 minutes.

(a) Calculate the oxide thickness x_{ox} after the steam oxidation step.

(b) Calculate the final oxide thickness x_{ox} after the dry O_2 growth is completed.

5.2 A p-tub CMOS process uses a 2-step boron diffusion into an n-type wafer that has a constant donor doping of 10^{15} [cm^{-3}]. The boron diffusion is described by the following diffusion schedule:

Predeposit: 30 minutes at 900 [°C],
Drive-in: 75 minutes at 1100 [°C].

(a) Calculate the depth x_j of the p-tub region after the diffusion.

(b) Plot the boron doping profile as a function of position x [μm] superposed on the n-type donor level of 10^{15} [cm^{-3}].

(c) Suppose that the drive-in is increased to 90 minutes. Find the p-tub depth (i.e., x_j) for this case.

5.3 The wafer in Problem 5.2 is subjected to additional heat treatment steps, which are summarized as follows:

30 minutes at 1000 [°C]
60 minutes at 950 [°C]
45 minutes at 1100 [°C].

Find the new location of the pn junction after these steps are

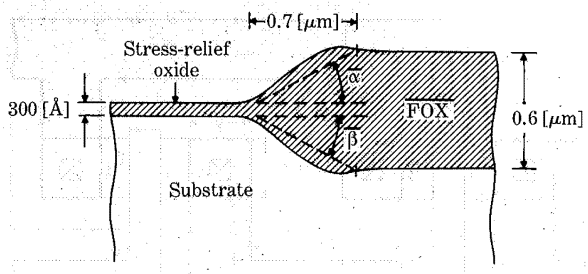


FIGURE P5.1

completed by using $\overline{Dt}$ in place of the drive-in (Dt) product. (Assume that the drive-in lasts 75 minutes as in the original problem statement.)

5.4 Phosphorus is implanted into silicon with an energy of 200 [keV], which results in a projected range of $R_p \simeq 0.25$ [μm] and a straggle of $\Delta R_p \simeq 0.08$ [μm]. The implant dose is 4×10^{14} [cm^{-3}], and the wafer doping is constant at a level of $N_a = 10^{15}$ [cm^{-3}] (p-type).
(a) Calculate the peak implant density N_p [cm^{-3}].
(b) What is the surface implanted doping level N_{ion} ($x = 0$)?
(c) Calculate the junction depth x_j where the pn junction is formed.

5.5 A LOCOS process yields the bird's beak cross section shown in Fig. P5.1. The field oxide is $X_{\text{FOX}} = 0.60$ [μm], and the stress-relief oxide has a thickness of 300 [Å]. The lateral encroachment region where the bird's beak is formed is 0.7 [μm]. Calculate the bird's beak angles α and β shown in the drawing.

5.6 A polysilicon layer is doped n-type during a self-aligned MOSFET process flow. The poly doping is approximated by $N_{d,\text{poly}} \simeq 10^{20}$ [cm^{-3}], and the electron mobility is $\mu_n \simeq 150$ [cm^2/V-sec]. The poly layer has a nominal thickness of 0.20 [μm].
(a) Determine the sheet resistance R_s [Ω] of the poly layer.
(b) Find the resistance R [Ω/μm] if the width of the line is 3 [μm].
(c) Find the percent increase in R [Ω/μm] if the thickness of the poly layer is reduced to 0.18 [μm].

5.7 A field oxide is grown to a thickness of $X_{\text{FOX}} = 0.60$ [μm] on a p-type substrate that has a doping density of $N_a = 10^{15}$ [cm^{-3}]. The FOX charge is $Q_{ss,\text{FOX}} = q(5 \times 10^{10})$ [C/cm^2] $\gg Q_{\text{FOX}}$.
(a) Calculate the value of $(V_{T0})_{\text{FOX}}$ assuming an n-type poly interconnect with a doping density of $N_{d,\text{poly}} = 10^{20}$ [cm^{-3}]. (Use the results of Problem 1.2 to calculate Φ_{fS}.).
(b) Find the dose D_I of the field acceptor ion implant needed to set $(V_{T0})_{\text{FOX}}$ to a value of +15 [V].

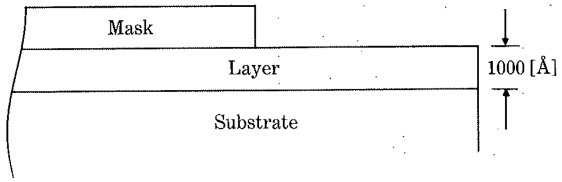


FIGURE P5.2

5.8 A wafer processing system using trench isolation results in the cross section of Fig. 5.30 with $X_{SOX} = 1500$ [Å], $X_{BOX} = 1700$ [Å], and $X_{TOX} = 4500$ [Å]. The trench itself is 1.4 [μm] wide and 5 [μm] deep.
(a) Calculate the poly-substrate capacitance per micron of trench.
(b) Calculate the interconnect-poly capacitance per micron of trench.

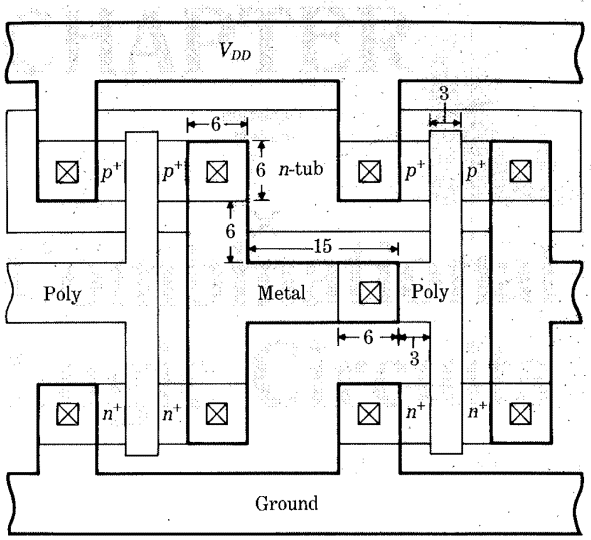
5.9 The MOSFET body bias coefficient is given by

$$\gamma = \frac{\sqrt{2q\varepsilon_{Si}N_a}}{C_{ox}}.$$

Suppose that the process gives nominal values of $x_{ox} = 500$ [Å] and $N_a = 10^{15}$ [cm^{-3}]. The oxide thickness is controlled to within 5%, while the acceptor doping N_a is known to have 8% variations.
(a) Calculate the percentage variation in γ.
(b) Find the minimum and maximum values of γ.

5.10 The layer shown in Fig. P5.2 is etched for a time $t_e = 1000$ [Å]/v_{ver} using different etching mechanisms. Assume that the masking layer is not affected.
(a) Provide scaled drawings of the etched layer cross section for degree of anisotropy values of $A = 0$, 0.25, 0.5, 0.75, and 1.0.
(b) Suppose that the etch time is increased to 1.5 t_e and that the substrate etch rates are identical to the etch rates for the layer. Sketch the profile if $A = 0.75$.

5.11 Cascaded CMOS inverters are shown in Fig. P5.3. A single stage of this layout was analyzed in Problem 4.10 of Chapter 4. Interconnect capacitances will now be added to the calculation. Assume interconnect levels of $C_{p\text{-}f} = 0.0576$ [fF/μm^2] and $C_{m\text{-}f} = 0.0345$ [fF/μm^2].
(a) Calculate the metal-field capacitance from the output of the first inverter to the metal-poly contact cut at the input to the second stage. Include only the metal-field lengths, i.e., ignore the regions where the metal overlaps with n^+, p^+, and poly.
(b) Find the input capacitance of the second stage as seen looking into the poly line. Then determine the sum ($C_{line} + G_G$).
(c) Compare the value of ($C_{line} + C_G$) to the value of C_{out} used in Problem 4.10. Does one contribution dominate?



All dimensions in [μm]

FIGURE P5.3

*(d) Prepare a SPICE file that models both the inverter capacitances and the interconnect contributions. Study the transient response and compare the values of t_{HL} and t_{LH} [ns].

5.12 A certain process gives an active area encroachment of 0.4 [μm] per side and a lateral doping distance of $L_o = 0.3$ [μm] (gate overlap). Give the mask layouts for MOSFETs that have final (W/L) values of 0.5, 1, and 4. Assume that the minimum resolution of the system is 3 [μm].

5.13 *Stick diagrams* are simplified drawings that show the placement and routing of IC layers. Figure P5.4 shows a sample base set of stick diagram notation where each material layer is represented by a specific line type. (An alternate approach is to use a specific color for each layer. In some instances, the color chosen for each layer roughly corresponds to the color seen for the layer on a wafer illuminated by white light. Otherwise, the choice of colors has become somewhat arbitrary.)

Chip layout patterns can be described using stick diagrams. These are useful for initial planning. In a self-aligned MOS technology, the gate poly layer is used to mask the n^+ pattern in making nMOS transistors. Thus, assuming that a threshold adjustment implant is provided, drawing a poly stick that crosses an n^+ stick gives an E-mode MOSFET. A D-mode transistor is obtained by adding a depletion implant.

(a) Construct the stick diagrams for the three types of nMOS inverters.

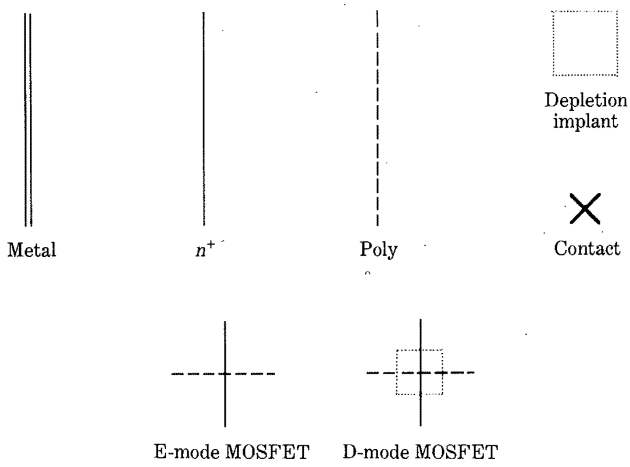


FIGURE P5.4

Then discuss the advantage of having a stick diagram vs. the information that is missing from the drawing.

(b) Create additional lines (or colors) to describe an n-tub CMOS process. Then construct the stick diagram for a CMOS inverter.

5.14 Consider a depletion load inverter in a process that exhibits threshold voltage variations of ΔV_{T0D} and ΔV_{T0L}.

(a) Starting with the design equation for β_R in eqn. (3.4-43), compute the change $\Delta\beta_R$ due to variations in V_{T0D} and V_{T0L}.

(b) Assume that $V_{OH} = 5$ [V], $\gamma = 0.37$ [V], $2|\phi_F| = 0.58$ [V], and

$$V_{T0D} = 0.9 \pm 0.1\,[\text{V}], \qquad V_{T0L} = -3.3 \pm 0.1\,[\text{V}]$$

give the threshold voltage ranges. A design value of $V_{OL} = 0.25$ [V] is set as a design specification. Use the expression derived in (a) to calculate the variation $\Delta\beta_R$. Then discuss the implications the analysis has on overall design acccuracy and chip yield.