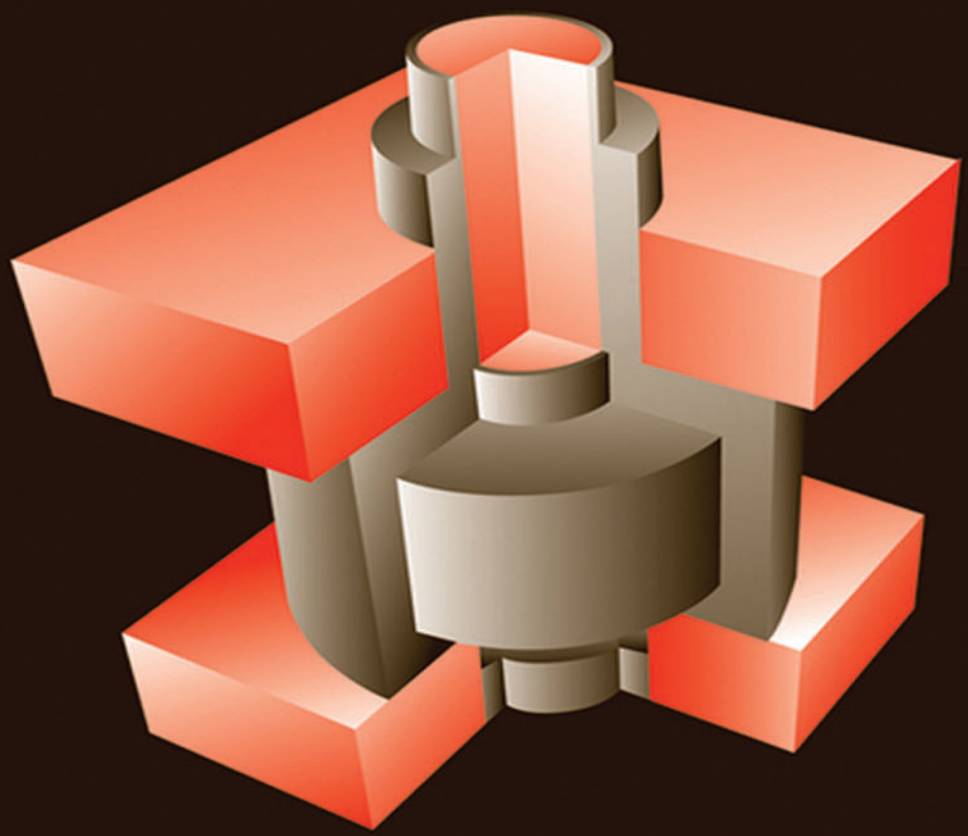
Semiconductor Devices

Physics and Technology

S. M. SZE | M. K. LEE



3RD EDITION

3RD EDITION

Semiconductor Devices

Physics and Technology

S. M. SZE
EtronTech Distinguished Chair Professor
College of Electrical and Computer Engineering
National Chiao Tung University
Hsinchu, Taiwan

M. K. LEE
Professor
Department of Electrical Engineering
National Sun Yat-sen University
Kaohsiung, Taiwan

JOHN WILEY & SONS, INC.

Acquisitions Editor *Dan Sayre*
Marketing Manager *Christopher Ruel*
Senior Editorial Assistant *Katie Singleton*
Editorial Program Assistant *Samantha Mendel*
Production Manager *Micheline Frederick*
Cover Designer *Wendy Lai*
Pre-press Service *Robots & Cupcakes*

This book was typeset in *Times Roman* by *the authors* and printed and bound by *Quad Graphics/Versailles.* The cover was printed by *Quad Graphics/Versailles.*

cover photo: © 2010 IEEE. Reprinted, with permission, from IEDM Technical Digest, S. Whang et. al, "Novel 3-dimensional Dual Control-gate with Surrounding Floating-gate (DC-SF) NAND flash cell for 1Tb file storage application."

The book is printed on acid-free paper.

ISBN 978-0470-53794-7

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

6.2 CMOS AND BiCMOS

Complementary MOS (CMOS) refers to a complementary *p*-channel and *n*-channel MOSFET pair. CMOS logic is the most popular technology utilized in present-day integrated circuit design. The main reasons for the success of CMOS are low power consumption and good noise immunity.

6.2.1 The CMOS Inverter

A CMOS inverter, which is the basic element of CMOS logic circuits, is shown in Fig. 12. In a CMOS inverter, the gates of the *p*- and *n*-channel transistors are connected and serve as the input node to the inverter. The drains of the two transistors are also connected and serve as the output node to the inverter. The source and substrate contacts of the *n*-channel MOSFET are grounded, whereas those of the *p*-channel MOSFET are connected to the power supply (V_{DD}). Note that both *p*-channel and *n*-channel MOSFETs are enhancement-type transistors. When the input voltage is low (e.g., $V_{in} = 0$, $V_{GSn} = 0 < V_{Tn}$), the *n*-channel MOSFET is off.* The *p*-channel MOSFET, however, is on, since $|V_{GSp}| \cong V_{DD} > |V_{Tp}|$ (V_{GSp} and V_{Tp} are negative). Consequently, the output node is charged to V_{DD} through the *p*-channel MOSFET. When the input voltage goes high so that the gate voltage equals V_{DD}, the *n*-channel MOSFET is turned on, since $V_{GSn} = V_{DD} > V_{Tn}$, and the *p*-channel MOSFET is turned off, since $|V_{GSp}| \cong 0 < |V_{Tp}|$. Therefore, the output node is discharged to ground through the *n*-channel MOSFET.

For a more detailed understanding of the operation of the CMOS inverter, we can plot the output characteristics of the transistors. This plot is given in Fig. 13, in which I_p and I_n are shown as a function of output voltage (V_{out}). I_p is the current of *p*-channel MOSFET in the direction from the source (connected to V_{DD}) to the drain (output node). I_n is the current of *n*-channel MOSFET in the direction from the drain (output node) to the source (connected to ground). Note that the increase in input voltage (V_{in}) tends to increase I_n but decrease I_p at fixed V_{out}. In steady state, however, I_n should be equal to I_p. For a given V_{in}, we can determine the corresponding V_{out} from the intercept of $I_n(V_{in})$ and $I_p(V_{in})$, as shown in Fig. 13. The V_{in}-V_{out} curve, as shown in Fig. 14, is called the transfer curve of the CMOS inverter.[4]

An important characteristic of the CMOS inverter is that when the output is in a steady logic state, i.e., $V_{out} = 0$ or V_{DD}, only one transistor is on. The current flow from the power supply to ground is thus very low and is equal to the leakage current of the off device. In fact, there is significant current conduction only during the short transient period when the two devices are temporarily on. Therefore the power consumption is very low in the static state compared with other types of logic circuits, such as *n*-channel MOSFETs, bipolar, etc.

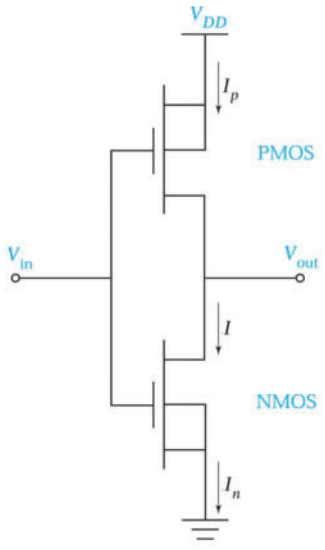


Fig. 12 **The CMOS inverter.**

* V_{GSn} and V_{GSp} are the voltage differences between the gate and the source for *n*- and *p*-channel MOSFETs, respectively.

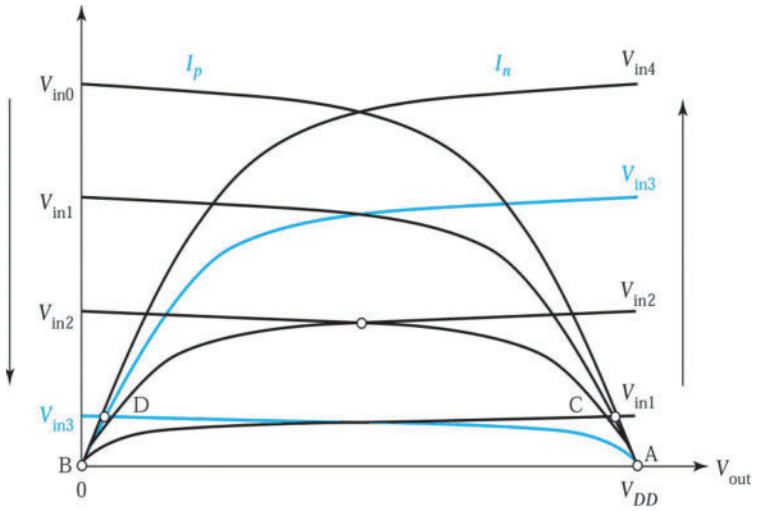


Fig. 13 **I_p and I_n as functions of V_{out}. The intercepts of I_p and I_n (circled) represent the steady-state operation points of the CMOS inverter.[4] The curves are labeled by the input voltages: $0 = V_{in0} < V_{in1} < V_{in2} < V_{in3} < V_{in4} = V_{DD}$.**

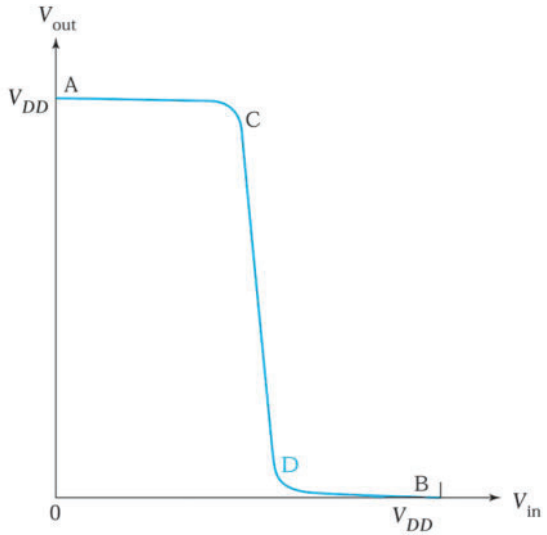


Fig. 14 **Transfer curve of a CMOS inverter.[4] Points labeled A, B, C, and D correspond to the points labeled in Fig. 13.**

6.2.2 Latch-up

In order to fabricate both *p*-channel and *n*-channel MOSFETs in the same chip for CMOS applications, extra doping and diffusion steps are needed to form the "well" or "tub" in the substrate. The doping type in the well is different from that of the surrounding substrate. Typical well types are the *p*-well, *n*-well, and twin well. Details of the well technology are given in Chapter 15. Figure 15 shows a cross-sectional view of a CMOS inverter fabricated using *p*-well technology. In this figure, the *p*-channel and *n*-channel MOSFETs are fabricated in the *n*-type Si substrate and the *p*-well region, respectively.

A major problem related to the well structure in CMOS circuits is the latch-up phenomenon. The cause of latch-up is the action of the parasitic *p*-*n*-*p*-*n* diode in the well structure. As shown in Fig. 15, the parasitic *p*-*n*-*p*-*n* diode consists of a lateral *p*-*n*-*p* and a vertical *n*-*p*-*n* bipolar transistor. The *p*-channel MOSFET's source, *n*-substrate, and *p*-well correspond to the emitter, base, and collector of the lateral *p*-*n*-*p* bipolar transistor,

respectively. The *n*-channel MOSFET's source, *p*-well, and *n*-substrate are the emitter, base, and collector of the vertical *n*-*p*-*n* bipolar transistor, respectively. The equivalent circuit of the parasitic components is illustrated in Fig. 16, where R_S and R_W are the series resistance in the substrate and the well, respectively. The base of each transistor is driven by the collector of the other to form a positive feedback loop. This configuration is similar to the thyristor discussed in Chapter 4. Latch-up is induced when the current gain product of the two bipolar transistors, $\alpha_{npn}\alpha_{pnp}$, is larger than 1. When latch-up occurs, a large current will flow from the power supply (V_{DD}) to the ground contact. This can interrupt normal circuit operation and even destroy the chip itself because of the high power dissipation required.

To avoid latch-up, the current gains of the parasitic bipolar transistors must be reduced. One method is to use gold doping or neutron irradiation to lower the minority carrier lifetimes. However, this approach is difficult to control. Besides, it also causes an increase of the leakage current. A deeper well structure or high-energy implantation to form retrograde wells can also reduce the current gain of the vertical bipolar transistor by raising the impurity concentration in the base. In the retrograde well, the peak of the well doping concentration is located within the substrate away from the surface.

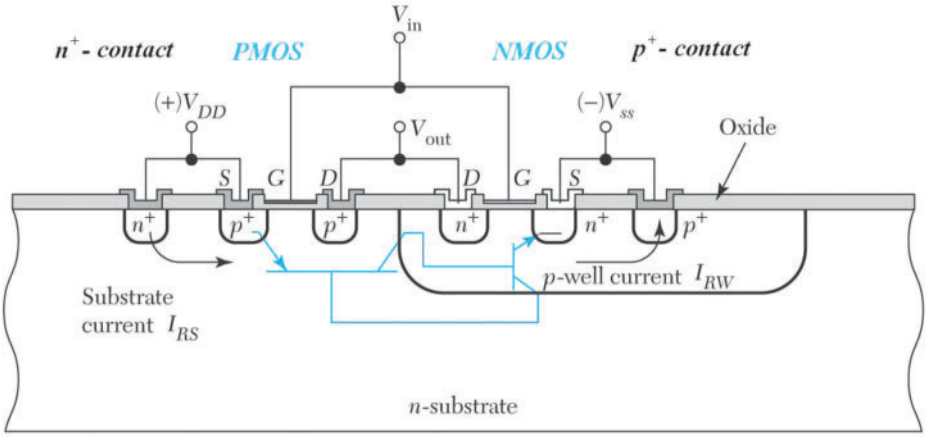


Fig. 15 **Cross section of a CMOS inverter fabricated with *p*-well technology.**

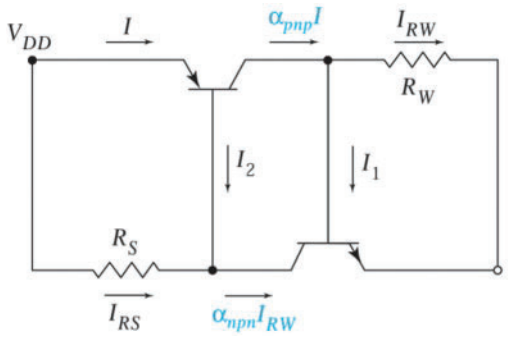


Fig. 16 **Equivalent circuit of the *p*-well structure shown in Fig. 15.**

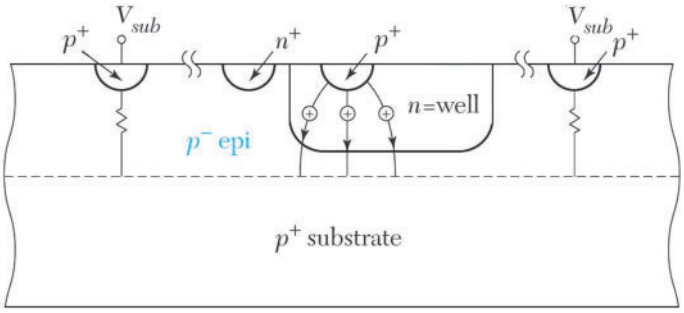


Fig. 17 **Prevention of latch-up with a heavily doped substrate.**[7]

Another way to reduce latch-up is to use a heavily doped substrate with devices fabricated on a lightly doped epitaxial layer, as shown[7] in Fig. 17. The heavily doped substrate provides a highly conductive path to collect the current. The current then is drained away through the surface contacts (V_{sub}).

Latch-up can also be avoided with the trench isolation scheme. A process for forming trench isolation is discussed in Chapter 15. This approach can eliminate latch-up because the *n*-channel and *p*-channel MOSFETs are physically isolated by the trench.

6.2.3 CMOS Image Sensor

For consumer imaging products such as digital cameras and video recorders, the CCD image sensor discussed in Section 5.4, Chapter 5, dominates the market. However, since the late 1990s[8] this huge market has been increasingly eroded by the CMOS image sensor fabricated by using standard CMOS processes.

In principle a CMOS image sensor, shown in Fig. 18,[9] has a very similar architecture to a semiconductor memory. It is composed of an array of identical pixels. Each pixel has a photodiode (a *p*-*n* junction photodiode[10]), that converts incident light into photocurrent, and an addressing transistor that acts as a switch, as shown in Fig. 19*a*. A Y-addressing or scan register is used to address the sensor line by line, by activating the in-pixel addressing transistor. An X-addressing or scan register is used to address the pixels on one line, one after another. Some of the readout circuits need to convert the photocurrent into electric charge or voltage and to read it off the array.

The working principle of a pixel is as follows: (1) at the beginning of an exposure the photodiode is reverse biased to a high voltage; (2) during the exposure time, impinging photons decrease the reverse voltage across the photodiode; (3) at the end of the exposure time the remaining voltage across the diode is measured, and the voltage drop from the original value is a measure of the number of photons falling on the photodiode during the exposure time; (4) the photodiode is reset to allow a new exposure cycle.

The most basic form of imaging array shown in Fig. 19b is called PPS (passive pixel sensor), where in each pixel a select transistor controls each photodetector. The advantage is that many cells in a row are accessed at the same time, as in a memory array, so the speed is higher than CCD whose readout is serial in nature. The penalty is larger size.

Many of the differences between CCD and CMOS image sensors arise from differences in their readout architectures. In a CCD (see Fig. 20, Chapter 5), charge is shifted out of the array via vertical and horizontal CCDs, converted into voltage via a simple follower amplifier, and then serially read out. In a CMOS image sensor, charge voltage signals are read out one row at a time in a manner similar to a random-access memory using row and column select circuits.

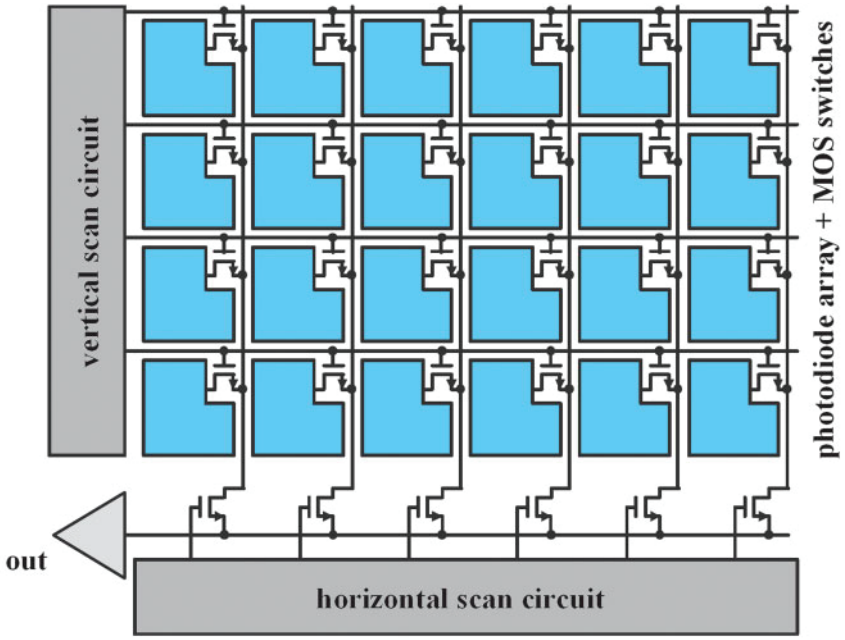


Fig. 18 **Architecture of a two-dimensional CMOS image sensor.**

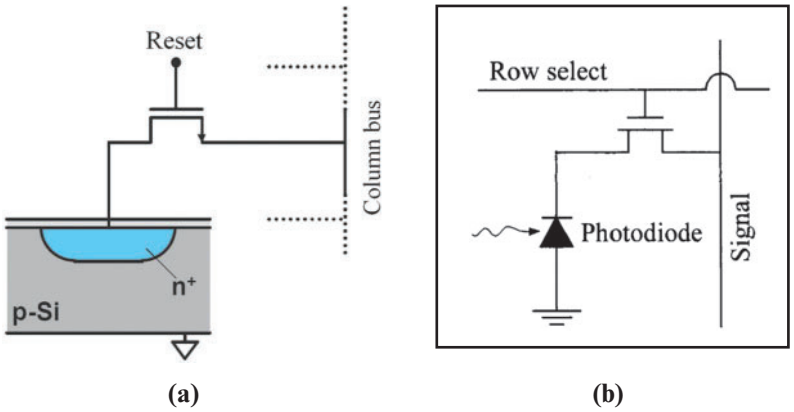


Fig. 19 **(*a*) Passive CMOS pixel based on a single in-pixel transistor. (*b*) PPS (passive pixel sensor) CMOS image sensors.**

The replacement of CCD by CMOS image sensors is growing due to the increasing integration of more functionality within each pixel discussed above, taking advantages of the conventional CMOS scaling and inexpensive technology. Moreover, the advantages of the CMOS image sensor include higher speed due to random-access capability, larger signal-to-noise ratio, lower power due to low vo1tage requirements, and low cost because of mainstream technology. Conversely, the CCD maintains some advantages such as small pixe1 size, low-light sensitivity, and high dynamic range. However, CCD requires a different process optimization, so CCD systems that include CMOS circuitry are naturally more expensive.

6.2.4 BiCMOS

CMOS has the advantages of low power dissipation and high device density that make it suitable for fabricating complex circuits. However, CMOS suffers from low drive capability compared with bipolar technology, which limits its circuit performance. BiCMOS is a technology that integrates both CMOS and bipolar device structures in the same chip. A BiCMOS circuit contains mostly CMOS devices, with a relatively small number of bipolar devices. The bipolar devices have better switching performance than their CMOS counterparts without consuming too much extra power. However, this performance enhancement is achieved at the expense of extra manufacturing complexity, longer fabrication time, and higher cost. The fabrication processes for BiCMOS are discussed in Chapter 15.

▶ 6.3 MOSFET ON INSULATOR

For certain applications, MOSFETs are fabricated on an insulating substrate rather than on a semiconductor substrate. The characteristics of these transistors are similar to those of an MOSFET. Usually, we call such devices thin film transistors (TFT) if the channel layer is an amorphous or polycrystalline silicon. If the channel layer is a monocrystalline silicon, we call it silicon-on-insulator (SOI).

6.3.1 Thin Film Transistor (TFT)

Hydrogenated amorphous silicon (a-Si:H) and polysilicon are the two most popular materials for TFT fabrication. They are usually deposited on an insulating substrate such as a glass, quartz, or Si substrate with a thin SiO_2 capping layer.

The a-Si:H TFT is an important device in electronic applications that require a large area, such as liquid crystal displays (LCD) and contact imaging sensors (CIS). The a-Si:H materials are usually deposited with a plasma-enhanced chemical vapor deposition (PECVD) system. Since the deposition temperature is low (typically 200° – 400°C), inexpensive substrate materials such as glass can be used. The role played by the hydrogen atoms contained in the a-Si:H is to passivate dangling bonds in the amorphous silicon matrix and thus reduce the defect density. Without hydrogen passivation, the gate voltage cannot adjust the Fermi level at the insulator and the a-Si interface, since the Fermi level is pinned by the large amount of defects.

The a-Si:H TFT is usually fabricated using the inverted staggered structure, as shown in Fig. 20. The inverted staggered structure is a bottom-gate scheme. A metal gate can be used since the post-process temperature is low (< 400°C). A dielectric layer such as silicon nitride or silicon dioxide, also deposited by PECVD, is often used as the gate dielectric. An undoped a-Si:H layer is subsequently deposited to form the channel. The source and drain of the TFT are formed with an in situ–doped n^+ a-Si:H layer complying with the requirement of low process temperature. A dielectric layer that serves as an etch- stop for patterning of n^+ a-Si:H is often used. Device characteristics of TFTs with the bottom-gate structure are usually better than those with the top-gate structure. This is because the a-Si:H channel could be damaged by plasma during PECVD gate-dielectric deposition of top-gated TFTs. In addition, the source/drain formation process is easier for the bottom-gate structure. A typical subthreshold characteristic of the a-Si:H TFT is shown in Fig. 21. Because of the amorphous matrix present in the channel material, its carrier mobility is usually very low (< 1 cm^2/V-s).

The polysilicon TFT uses a thin polysilicon as the channel layer. Polysilicon consists of many Si grains. Within the grains are the monocrystalline Si lattices. The orientations of two side-by-side grains are, however, different from each other. The interface between the two grains is called the grain boundary. Polysilicon TFT exhibits much higher carrier mobility and thus better drive capability than a-Si:H TFT because of higher crystallinity. Carrier mobility of these devices typically ranges from 10 to several hundred cm^2/V-s, depending on the grain size and process conditions. Polysilicon is usually deposited with low-pressure CVD (LPCVD). The grain size of polysilicon is an important factor in determining TFT performance, since the carrier mobility generally decreases with decreasing grain size. This is mainly because of the large number of defects contained in the grain boundaries that impede the transport of carriers.

where $\mathscr{E}$ is the electric field inside the photocoductor and υ_d is the carrier drift velocity. Substituting n in Eq. 4 into Eq. 5 gives

$$I_p \quad q\left(\eta \frac{P_{opt}}{h\nu}\right)\cdot\left(\frac{\mu_n \tau \mathscr{E}}{L}\right). \tag{6}$$

If we define the primary photocurrent as

$$I_{ph} \quad q\left(\eta \frac{P_{opt}}{h\nu}\right), \tag{7}$$

the photocurrent gain from Eq. 6 is

$$\boxed{\text{Gain} \equiv \frac{I_p}{I_{ph}} \quad \frac{\mu_n \tau \mathscr{E}}{L} \quad \frac{\tau}{t_r},} \tag{8}$$

where $t_r \equiv L/\upsilon_d = L/\mu_n \mathscr{E}$ is the carrier transit time. The gain depends on the ratio of carrier lifetime to the transit time.

▶ EXAMPLE 1

Calculate the photocurrent and gain when 5×10^{12} photons/s are arriving at the surface of a photoconductor of $\eta = 0.8$. The minority carrier lifetime is 0.5 ns, and the device has $\mu_n = 2500$ cm^2/ V-s, $\mathscr{E} = 5000$ V/cm, and $L = 10$ μm.

SOLUTION From Eq. 6,

$$I_p = q\left(0.8\times 5\times 10^{12} \text{ photons/s}\right)\cdot\left(\frac{2500 \text{ cm}^2/\text{V-s}\cdot 5\times 10^{-10}\text{ s}\cdot 5000 \text{ V/cm}}{10\times 10^{-4}\text{ cm}}\right)$$

$$= 4\times 10^{-6}\text{ A} = 4 \text{ μA}$$

and from Eq. 8,

$$\text{Gain} \quad \frac{\mu_n \tau \mathscr{E}}{L} \quad \frac{2500\cdot 5\times 10^{-10}\cdot 5000}{10\times 10^{-4}} \quad 6.25. \qquad \blacktriangleleft$$

For a sample with long minority-carrier lifetime and short electrode spacing, the gain can be substantially greater than unity. Gains as high as 10^6 can be obtained from some photoconductors. The response time of a photoconductor is determined by the transit time t_r. To achieve short transit time requires that we use small electrode spacing and a high electric field. The response times of photoconductors cover a wide range, from 10^{-3} to 10^{-10} seconds. They are extensively used for infrared detection, especially for wavelengths greater than a few micrometers.

10.1.2 Photodiode

A photodiode is basically a *p-n* junction operated under reverse bias. The space-charge and the electric-field distributions are similar to those in Fig. 6 in Chapter 3 except under reverse bias. Note that the electric-field distribution is nonuniform and the maximum field is at the junction. When an optical signal penetrates into the depletion region of the photodiode, the electric field in the depletion region serves to separate the photogenerated EHPs (electron-hole pairs) and an electric current, called photocurrent I_p, flows in the external circuit. The photogenerated holes drift in the depletion region, diffuse into the neutral *p* region, and then combine with

electrons entered from the negative electrode. Similarly, photogenerated electrons drift in the opposite direction. When an optical signal penetrates within a diffusion length outside the depletion region, the photogenerated carriers will diffuse into the depletion region and drift across the depletion region to the other side. These neutral regions can be regarded as resisitive extensions of electrodes to the depletion region. The photocurrent depends on the number of photogenerated EHPs and the drift velocities of the carriers. It should be noted that the photocurrent in the external circuit is due only to the flow of electrons, even though there is electron and hole drift in the depletion region.

For high-frequency operation, the depletion region must be kept thin to reduce the transit time. On the other hand, to increase the quantum efficiency the depletion layer must be sufficiently thick to allow a large fraction of the incident light to be absorbed. Thus, there is a trade-off between the response speed and quantum efficiency.

Quantum Efficiency

The quantum efficiency, as mentioned above, is the number of EHPs generated for each incident photon:

$$\eta = \left(\frac{I_p}{q}\right) \cdot \left(\frac{P_{opt}}{h\nu}\right)^{-1}, \tag{9}$$

where I_p is the photogenerated current from the absorption of incident optical power P_{opt} at a wavelength λ (corresponding to a photon energy $h\nu$) and is known more specifically as the external quantum efficiency. The internal quantum efficiency is defined as the photogenerated number of EHPs per absorbed photon. One of the key factors that determine η is the absorption coefficient α (Fig. 5, Chapter 9). Since α is a strong function of the wavelength, the wavelength range in which appreciable photocurrent can be generated is limited. The long-wavelength cutoff λ_c is established by the bandgap (Eq. 9, Chapter 9) and is about 1.8 μm for germanium and 1.1 μm for silicon. For wavelengths longer than λ_c, the values of α are too small to give appreciable band-to-band absorption. For wavelengths much shorter than λ_c, the values of α are too large (~10^5 cm^{-1}), and hence the radiation is mostly absorbed very near the surface where recombination time is short. Therefore, the photocarriers can recombine before they can be collected in the depletion region of *p-n* junction.

The photogenerated carriers in the depletion region may disappear by recombination or by trapping without contributing to the photocurrent. The quantum efficiency is always less than unity. It depends on the absorption coefficient and the device structure. The quantum efficiency can be increased by reducing the surface reflection on the device to increase the absorption in the depletion region, and by preventing the recombination or trapping of carriers through improving material and device quality.

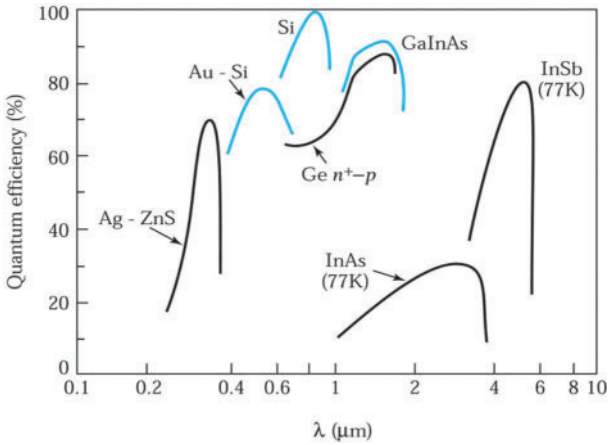


Fig. 2 Quantum efficiency versus wavelength for various photodetectors.[1,2]

CHAPTER 12

Film Formation

To fabricate discrete devices and integrated circuits, we use many different kinds of thin films. We can classify thin films into four groups: thermal oxides, dielectric layers, polycrystalline silicon, and metal films. Figure 1 shows a schematic view of a conventional silicon *n*-channel MOSFET (metal-oxide-semiconductor field-effect transistor) that uses all four groups of films. The first important thin film from the thermal oxide group is the gate-oxide layer, under which a conducting channel can be formed between the source and the drain. A related layer is the field oxide, which provides isolation from other devices. Both gate and field oxides generally are grown by a thermal oxidation process because only thermal oxidation can provide the highest-quality oxides having the lowest interface trap densities.

Dielectric layers such as silicon dioxide and silicon nitride are used for insulation between conducting layers, for diffusion and ion implantation masks, for capping doped films to prevent the loss of dopants, and for passivation to protect devices from impurities, moisture, and scratches. Polycrystalline silicon, usually referred to as polysilicon, is used as a gate electrode material in MOS devices, a conductive material for multilevel metallization, and a contact material for devices with shallow junctions. Metal films such as copper and silicides are used to form low-resistance interconnections, ohmic contacts, and rectifying metal-semiconductor barriers.

Specifically, we cover the following topics:

- The current density equation and its drift and diffusion components.
- The thermal oxidation process to form silicon dioxide (SiO_2).
- Chemical –vapor deposition techniques to form dielectrics and polysilicon films.
- Metallization and related global planarization.
- Atomic layer deposition to form thin films of the order of a monolayer.
- Characteristics of these thin films and their compatibility with integrated-circuit processing.

12.1 THERMAL OXIDATION

Semiconductors can be oxidized by various methods. These include thermal oxidation, electrochemical anodization, and plasma reaction. Among these methods, thermal oxidation is by far the most important for silicon devices. It is the key process in modern silicon integrated-circuit technology. For gallium arsenide, however, thermal oxidation results in generally nonstoichiometric films. The oxides provide poor electrical insulation and

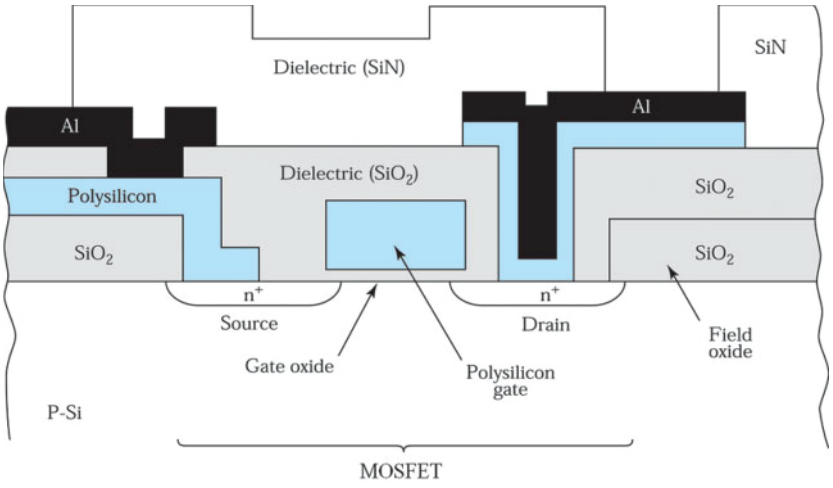


Fig. 1 Schematic cross section of a metal-oxide-semiconductor field-effect transistor (MOSFET).

semiconductor surface protection; hence, these oxides are rarely used in gallium arsenide technology. Consequently, in this section we concentrate on the thermal oxidation of silicon.

The basic thermal oxidation setup is shown[1] in Fig. 2. The reactor consists of a resistance-heated furnace, a cylindrical fused-quartz tube containing the silicon wafers held vertically in a slotted quartz boat, and a source of either pure dry oxygen or pure water vapor. The loading end of the furnace tube protudes into a vertical flow hood where a filtered flow of air is maintained. Flow is directed as shown by the arrow in Fig. 2. The hood reduces dust and particulate matters in the air surrounding the wafers and minimizes contamination during wafer loading. The oxidation temperature is generally in the range of 900°-1200 °C and the typical gas flow rate is about 1 liter/min. The oxidation system uses microprocessors to regulate the gas flow sequence, to control the automatic insertion and removal of silicon wafers, to ramp the temperature up (i.e., to increase the furnace temperature linearly) from a low temperature to the oxidation temperature so that the wafers will not warp due to sudden temperature change, to maintain the oxidation temperature to within ±1°C, and to ramp the temperature down when oxidation is completed.

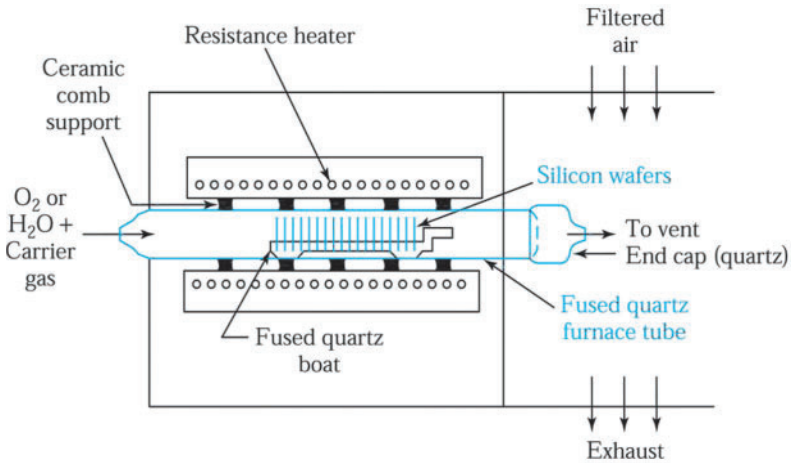


Fig. 2 Schematic cross section of a resistance-heated oxidation furnace.

12.1.1 Kinetics of Growth

The following chemical reactions describe the thermal oxidation of silicon in oxygen or water vapor:

$$\mathrm{Si(solid) + O_2\,(gas) \rightarrow SiO_2\,(solid)}, \tag{1}$$

$$\mathrm{Si(solid) + 2H_2O\,(gas) \rightarrow SiO_2\,(solid) + 2H_2\,(gas)}. \tag{2}$$

The silicon-silicon dioxide interface moves into the silicon during the oxidation process. This creates a fresh interface region with surface contamination on the original silicon ending up on the oxide surface. The densities and molecular weights of silicon and silicon dioxide are used in the following example to show that growing an oxide of thickness x consumes a layer of silicon $0.44x$ thick (Fig. 3).

▶ EXAMPLE 1

If a silicon oxide layer of thickness x is grown by thermal oxidation, what is the thickness of silicon being consumed? The molecular weight of Si is 28.9 g/mol, and the density of Si is 2.33 g/cm^3. The corresponding values for SiO_2 are 60.08 g/mol and 2.21 g/cm^3.

SOLUTION The volume of 1 mol of silicon is

$$\frac{\text{Molecular weight of Si}}{\text{Density of Si}} = \frac{28.9\ \text{g/mole}}{2.33\ \text{g/cm}^3} = 12.06\ \text{cm}^3\ \text{/ mol}.$$

The volume of 1 mol of silicon dioxide is

$$\frac{\text{Molecular weight of SiO}_2}{\text{Density of SiO}_2} = \frac{60.08\ \text{g/mol}}{2.21\ \text{g/cm}^3} = 27.18\ \text{cm}^3\ \text{/ mol}.$$

Since 1 mol of silicon is converted to 1 mol of silicon dioxide,

$$\frac{\text{Thickness of Si} \times \text{area}}{\text{Thickness of SiO}_2 \times \text{area}} = \frac{\text{volume of 1 mol of Si}}{\text{volume of 1 mol of SiO}_2},$$

$$\boxed{\frac{\text{Thickness of Si}}{\text{Thickness of SiO}_2} = \frac{12.06}{27.18} = 0.44,}$$

Thickness of silicon = 0.44 (thickness of SiO_2).
For example, to grow a silicon dioxide layer of 100 nm, a layer of 44 nm of silicon is consumed. ◀

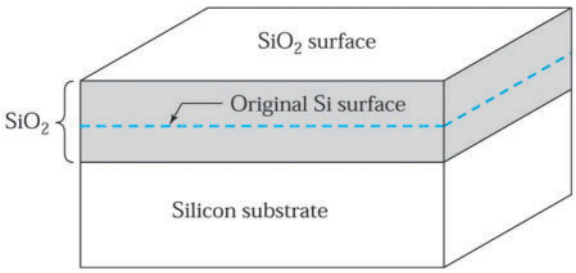


Fig. 3 **Growth of silicon dioxide by thermal oxidation.**

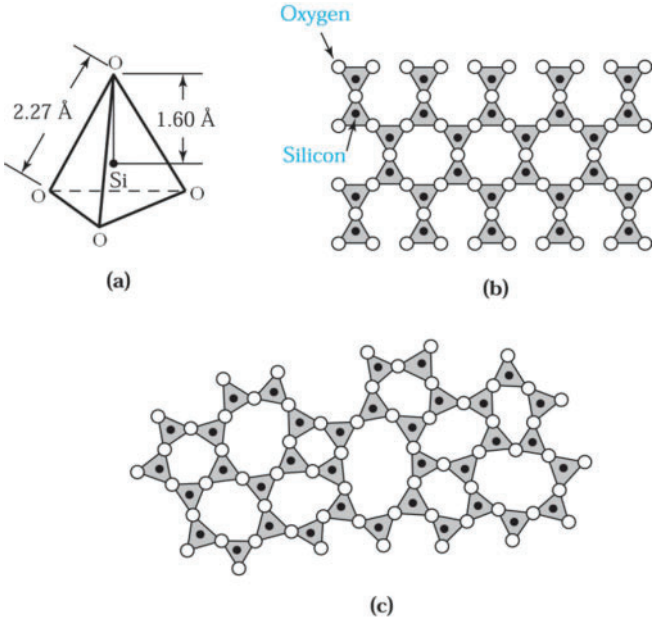


Fig. 4 (*a*) Basic structural unit of silicon dioxide. (*b*) Two-dimensional representation of a quartz crystal lattice. (*c*) Two-dimensional representation of the amorphous structure of silicon dioxide.[1]

The basic structural unit of thermally grown silicon dioxide is a silicon atom surrounded tetrahedrally by four oxygen atoms, as illustrated[1] in Fig. 4*a*. The silicon-to-oxygen internuclear distance is 1.6 Å, and the oxygen-to-oxygen internuclear distance is 2.27 Å. These tetrahedra are joined together at their corners by oxygen bridges in a variety of ways to form the various phases or structures of silicon dioxide (also called silica). Silica has several crystalline structures (e.g., quartz) and an amorphous structure. When silicon is thermally oxidized, the silicon dioxide structure is amorphous. Typically amorphous silica has a density of 2.21 g/cm^3 compared with 2.65 g/cm^3 for quartz.

The basic difference between the crystalline and amorphous structures is that the former is a periodic structure, extending over many molecules, whereas the latter has no periodic structure at all. Figure 4*b* is a two-dimensional schematic diagram of a quartz crystalline structure made up of rings with six silicon atoms. Figure 4*c* is a two-dimensional schematic diagram of an amorphous structure for comparison. In the amorphous structure there is still a tendency to form characteristic rings with six silicon atoms. Note that the amorphous structure in Fig. 4*c* is quite open because only 43% of the space is occupied by silicon dioxide molecules. The relatively open structure accounts for the lower density and allows a variety of impurities (such as sodium) to enter and diffuse readily through the silicon dioxide layer.

The kinetics of thermal oxidation of silicon can be studied using the simple model illustrated[2] in Fig. 5. A silicon slice contacts the oxidizing species (oxygen or water vapor), resulting in a surface concentration of C_0 molecules/cm^3 for these species. The magnitude of C_0 equals the equilibrium bulk concentration of the species at the oxidation temperature. The equilibrium concentration generally is proportional to the partial pressure of the oxidant adjacent to the oxide surface. At 1000 °C and a pressure of 1 atm, the concentration C_0 is 5.2×10^{16} molecules/cm^3 for dry oxygen and 3×10^{19} molecules/cm^3 for water vapor.

The oxidizing species diffuses through the silicon dioxide layer, resulting in a concentration C_s at the surface of silicon. The flux F_1 can be written as

$$F_1 = D\frac{dC}{dx} \cong \frac{D(C_0 - C_s)}{x}, \tag{3}$$

where D is the diffusion coefficient of the oxidizing species and x is the thickness of the oxide layer already present.

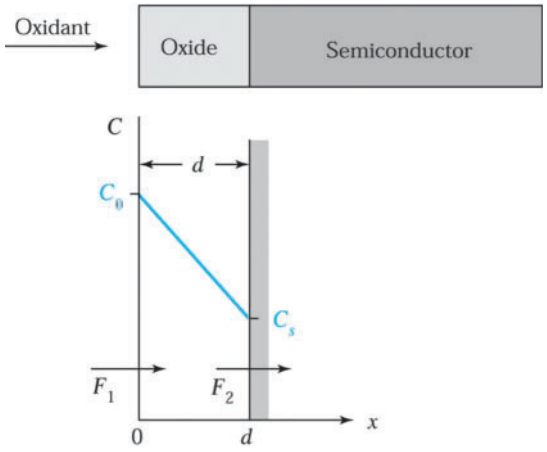


Fig. 5 **Basic model for the thermal oxidation of silicon.**[2]

At the silicon surface, the oxidizing species reacts chemically with silicon. Assuming the rate of reaction is proportional to the concentration of the species at the silicon surface, the flux F_2 is given by

$$F_2 = \kappa C_s, \tag{4}$$

where κ is the surface reaction rate constant for oxidation. At the steady state, $F_1 = F_2 = F$. Combining Eqs. 3 and 4 gives

$$F = \frac{DC_0}{x + (D/\kappa)}. \tag{5}$$

The reaction of the oxidizing species with silicon forms silicon dioxide. Let C_1 be the number of molecules of the oxidizing species in a unit volume of the oxide. There are 2.2×10^{22} silicon dioxide molecules/cm^3 in the oxide, and we add one oxygen molecule (O_2) to each silicon dioxide molecule, whereas we add two water molecules (H_2O) to each silicon dioxide molecule. Therefore, C_1 for oxidation in dry oxygen is 2.2×10^{22} cm^{-3}, and for oxidation in water vapor it is twice this number (4.4×10^{22} cm^{-3}). Thus, the growth rate of the oxide layer thickness is given by

$$\frac{dx}{dt} = \frac{F}{C_1} = \frac{DC_0/C_1}{x + (D/\kappa)}. \tag{6}$$

We can solve this differential equation subject to the initial condition x(0) = d_0, where d_0 is the initial oxide thickness; d_0 can also be regarded as the thickness of oxide layer grown in an earlier oxidation step. Solving Eq. 6 yields the general relationship for the oxidation of silicon:

$$x^2 + \frac{2D}{\kappa}x = \frac{2DC_0}{C_1}(t + \tau), \tag{7}$$

where $\tau \equiv (d_0^2 + 2Dd_0/\kappa)C_1/2DC_0$, which represents a time coordinate shift to take into account the initial oxide layer d_0.

The oxide thickness after an oxidizing time t is given by

$$x = \frac{D}{k}\left[\sqrt{1 + \frac{2C_0\kappa^2(t+\tau)}{DC_1}} - 1\right]. \tag{8}$$

For small values of t, Eq. 8 reduces to

$$x \cong \frac{C_0 \kappa}{C_1}(t+\tau), \tag{9}$$

and for larger values of t, it reduces to

$$x \cong \sqrt{\frac{2DC_0}{C_1}(t+\tau)}. \tag{10}$$

During the early stages of oxide growth, when the surface reaction is the rate-limiting factor, the oxide thickness varies linearly with time. As the oxide layer becomes thicker, the oxidant must diffuse through the oxide layer to react at the silicon-silicon dioxide interface and the reaction becomes diffusion limited. The oxide growth then becomes proportional to the square root of the oxidizing time, which results in a parabolic growth rate.

Equation 7 is often written in a more compact form:

$$\boxed{x^2 + Ax = B(t+\tau).} \tag{11}$$

where $A = 2D/\kappa$, $B = 2DC_0/C_1$ and $B/A = \kappa C_0/C_1$. Using this form, Eqs. 9 and 10 can be written as

$$\boxed{x = \frac{B}{A}(t+\tau)} \tag{12}$$

for the linear region and as

$$\boxed{x^2 = B(t+\tau).} \tag{13}$$

for the parabolic region. For this reason, the term B/A is referred to as the linear rate constant and B as the parabolic rate constant. Experimentally measured results agree with the predictions of this model over a wide range of oxidation conditions. For wet oxidation, the initial oxide thickness d_0 is very small, or $\tau \cong 0$. However, for dry oxidation, the extrapolated value of d_0 at $t = 0$ is about 20 nm.

The temperature dependence of the linear rate constant B/A is shown in Fig. 6 for both dry and wet oxidation and for (111)- and (l00)-oriented silicon wafers.[2] The linear rate constant varies as exp $(-E_a/kT)$, where the activation energy E_a is about 2 eV for both dry and wet oxidation. This agrees closely with the energy required to break silicon-silicon bonds, 1.83 eV/molecule. Under a given oxidation condition, the linear rate constant depends on crystal orientation. This is because the rate constant is related to the rate of incorporation of oxygen atoms into the silicon. The rate depends on the surface bond structure of silicon atoms, making it orientation dependent. Because the density of available bonds on the (111)-plane is higher than that on the (100)-plane, the linear rate constant for (111)-silicon is larger.

Figure 7 shows the temperature dependence of the parabolic rate constant B, which can also be described by $\exp(-E_a/kT)$. The activation energy E_a is 1.24 eV for dry oxidation. The comparable activation energy for oxygen diffusion in fused silica is 1.18 eV. The corresponding value for wet oxidation, 0.71 eV, compares favorably with the value of 0.79 eV for the activation energy of diffusion of water in fused silica. The parabolic rate constant is independent of crystal orientation. This independence is expected because it is a measure of the diffusion process of the oxidizing species through a random network layer of amorphous silica.

Although oxides grown in dry oxygen have the best electrical properties, considerably more time is required to grow the same oxide thickness at a given temperature in dry oxygen than in water vapor. For relatively thin oxides such as the gate oxide in a MOSFET (typically ≤ 20 nm), dry oxidation is used.

However, for thicker oxides such as field oxides (≥ 20 nm) in MOS integrated circuits and for bipolar devices, oxidation in water vapor (or steam) is used to provide both adequate isolation and passivation.

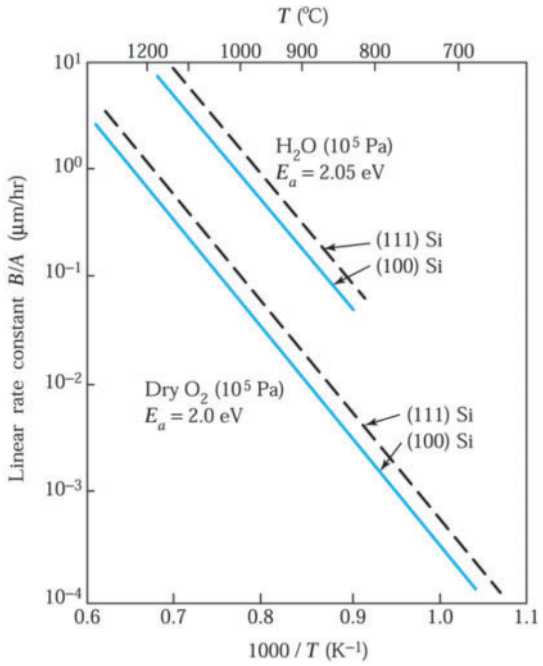


Fig. 6 **Linear rate constant versus temperature.**[2]

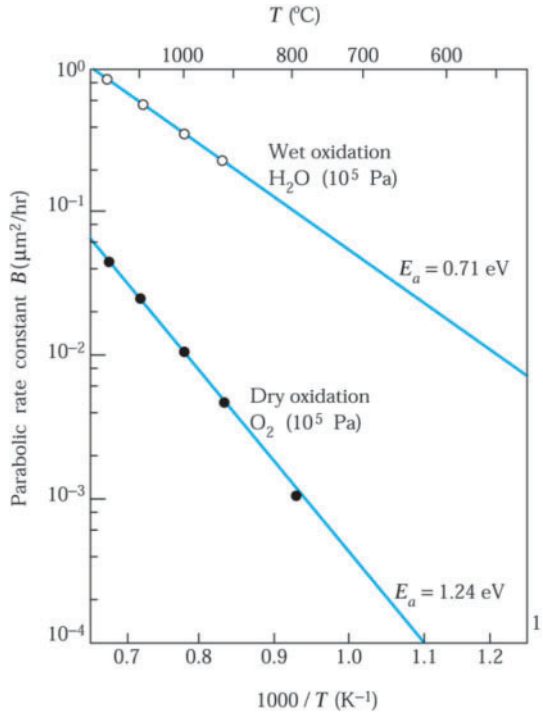


Fig. 7 **Parabolic rate constant versus temperature.**[2]

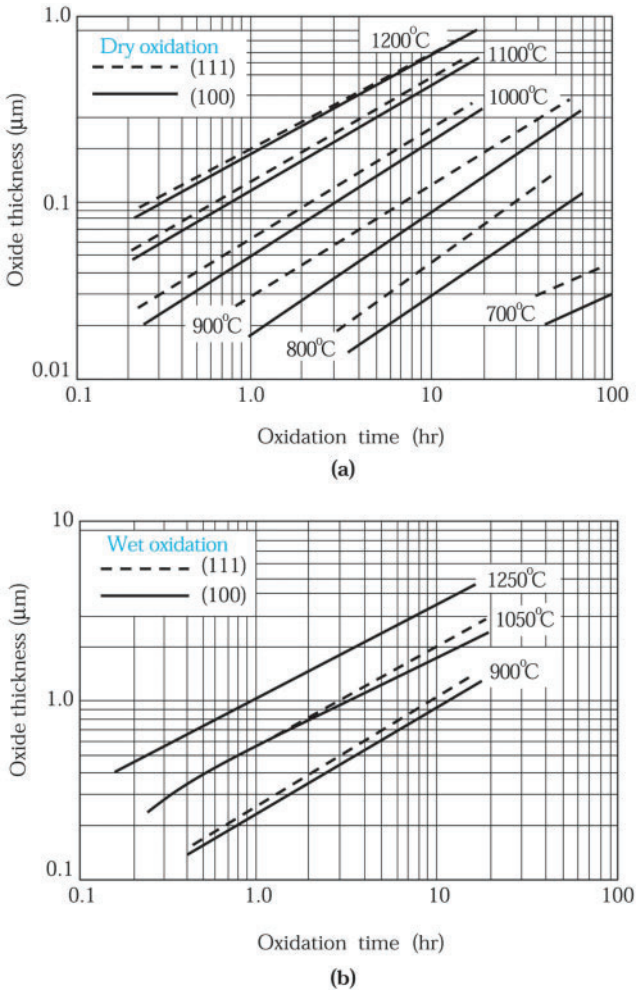


Fig. 8 Experimental results for silicon dioxide thickness as a function of reaction time and temperature for two substrate orientations. (*a*) Growth in dry oxygen. (*b*) Growth in steam.[3]

Figure 8 shows experimental results for silicon dioxide thickness as a function of reaction time and temperature for two substrate orientations.[3] Under a given oxidation condition, the oxide thickness grown on a (111)-substrate is larger than that grown on a (100)-substrate because of the larger linear rate constant of the (111)-orientation. Note that for a give temperature and time, the oxide film obtained using wet oxidation is about 5–10 times thicker than that obtained using dry oxidation.

▶ EXAMPLE 2

Using Fig. 8, determine the thickness of an SiO_2 layer grown on a (100) bare Si wafer in the following three sequential steps: (*a*) 60 min., 1200 °C, dry O_2, (*b*) 18 min., 900 °C, stream, (*c*) 30 min., 1050 °C, stream.

SOLUTION

(*a*) Since we are beginning with a bare silicon wafer, we can use Fig. 8*a* directly. We find a value of 0.18 μm or 180 nm.

(*b*) Using 0.18 μm as a starting point on Fig. 8*b*, we find that we have grown the equivalent of 0.7 hr or 42 min. We add another 18 min, bringing the total time to 60 min. Figure 8*b* shows a total oxide thickness of 0.22 μm.
(*c*) Using 0.22 μm as a starting point on Fig. 8*b*, we find that we have grown the equivalent of 15 min. We add another 30 min, bringing the total time to 45 min. Figure 8*b* shows a total oxide thickness of 0.48 μm. ◀

12.1.2 Thin Oxide Growth

Relatively slow growth rates must be used to grow thin oxide films of precise thicknesses reproducibly. Various approaches to achieving such slower growth rates have been reported. The mainstream approach for gate oxides 10–15 nm thick is to grow the oxide film at atmospheric pressure and lower temperatures (800°–900°C). With this approach, processing using modern *vertical* oxidation furnaces can grow reproducible, high-quality 10 nm oxides to within 0.1 nm across the wafer.

We noted earlier that for dry oxidation, there is an apparently rapid oxidation that gives rise to an initial oxide thickness d_0 of about 20 nm. Therefore, the simple model presented in Section 12.1.1 is not valid for dry oxidation with oxide thickness ≤ 20 nm. For ultralarge-scale integration (ULSI), the ability to grow thin (5–20 nm), uniform, high-quality reproducible gate oxides has become increasingly important. We briefly consider the growth mechanisms of such thin oxides.

In the early stage of growth in dry oxidation, there is a large compressive stress in the oxide layer that reduces the oxygen diffusion coefficient in the oxide. As the oxide becomes thicker, the stress will be reduced due to the viscous flow of silica and the diffusion coefficient will approach its stress-free value. Therefore, for thin oxides, the value of D/κ may be sufficiently small that we can neglect the term Ax in Eq. 11 and obtain

$$x^2 - d_0^2 = Bt, \tag{14}$$

where d_0 is equal to $\sqrt{2DC_0\tau / C_1}$, which is the initial oxide thickness when time is extrapolated to zero, and B is the parabolic rate constant defined previously. We therefore expect the initial growth in dry oxidation to follow a parabolic form.

▶ 12.2 CHEMICAL VAPOR DEPOSITION OF DIELECTRICS

Deposited dielectric films are used mainly for insulation and passivation of discrete devices and integrated circuits. Considerations in selecting a deposition process are the substrate temperature, the deposition rate and film uniformity, the morphology, the electrical and mechanical properties, and the chemical composition of the dielectric films.

12.2.1 Chemical Vapor Deposition

Chemical vapor deposition (CVD) is the most useful method for the deposition of a wide variety of thin films in semiconductor device fabrication. CVD is used to deposit, for example, polysilicon for gate conductor, silica glass, doped silica glass such as borophosphosilicate glass (BPSG) and phosphosilicate glass (PSG), silicon nitride for dielectric films, and tungsten, tungsten silicide, and titanium nitride for conducting films. Other emerging dielectrics such as high-dielectric-constant materials (e.g., hafnium silicate), low-dielectric-constant materials (e.g., carbon-doped silicate glass), and conductors (e.g., copper barrier/tantalum nitride, copper, ruthenium) can also be deposited by CVD.

There are three commonly used deposition methods: atmospheric-pressure CVD, low-pressure CVD (LPCVD), and plasma-enhanced chemical vapor deposition (PECVD, or plasma deposition). The reactor for atmospheric-pressure CVD is similar to the one shown in Fig. 2, except that different gases are used at the gas inlet. LPCVD is a CVD process operated at subatmospheric pressures. Reduced pressures can reduce unwanted gas-phase reactions and improve film uniformity across the wafer. However, it suffers from low deposition

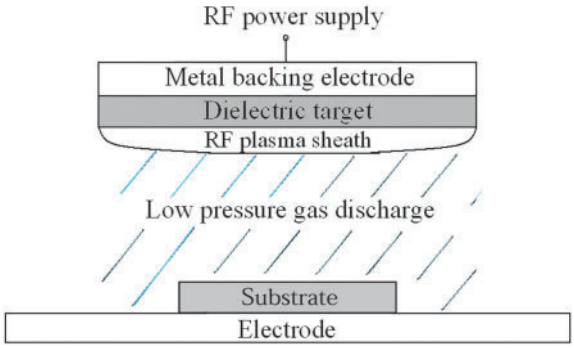


Fig. 21 **Schematic diagram of RF sputtering.**

RF sputtering

Rf (radio-frequency, typically 13.56 MHz, a frequency chosen because of its non-interference with radio-transmitted signals) sputtering is usually used in cases of dielectric materials, such as the high-k dielectrics. Figure 21 shows the standard rf sputtering system. Its has several advantages: (a) its ability to sputter dielectrics as well as metals, (b) its ability to operate in the bias-sputtering mode, and (c) its ability to permit sputter-etching of substrates prior to deposition. When a time-varying potential is applied to a metal plate behind the dielectric target in rf sputtering, another time-varying potential is developed on the opposite target surface through the impedance of the target. Once the gas is broken down by the acceleration of stray electrons from the electric field to start a discharge, the current can flow from the plasma to the target surface. Since electrons are more mobile than positive ions, more electrons are attracted to the front surface of the target during the positive half cycle than are positive ions in the negative half cycle. Therefore, the current is larger in the positive cycle than that in the negative cycle, as in a diode. The resultant electron current causes the target surface to acquire an increasingly negative bias voltage during successive cycles until the negative average dc voltage is sufficiently high to retard the electrons' arrival, so that the net charge arriving at the target surface is zero.

Since the target potential is negative with respect to the plasma, electrons are forced away from the surface, yielding an ion sheath that is visible as a dark space (because there is no optical emission from the recombination of electrons and ions) near the target surface. Positive ions in the sheath are accelerated toward the target by the negative potential. To prevent accumulation of excessive positive ions at the target surface, the frequency of the applied voltage must be high. The frequency must be at least 10^6 Hz for any appreciable sputtering to occur. Below this frequency, the average energy of the ions is reduced significantly as a result of positive ions accumulating on the target.

RF-sputter etching is the reverse of the sputtering process, and is also known as back sputtering, reverse sputtering, ion etching, or sputter cleaning. The normal rf power flow is electrically reversed; the substrate has a negative average dc voltage and an anode takes the place of the target. RF-sputter etching is used to clean substrates prior to sputtering a film on them, or to make patterns on substrates.

Bias-sputtering is the bombardment by energetic positive ions of a growing film that has a negative bias. This technology can remove impurities on the growing film. Usually, it is used for substrate surface cleaning before dielectric film deposition.

12.5.2 CVD Metal Deposition

CVD is attractive for metallization because it offers coatings that are conformal, has good step coverage, and can coat a large number of wafers at a time. The basic CVD setup is the same as that used for the deposition of dielectrics and polysilicon (see Fig. 9*a*). Low-pressure CVD (LPCVD) is capable of producing conformal step coverage over a wider range of topographical profiles, often with lower electrical resistivity than that from PVD.

One of the major new applications of CVD metal deposition for integrated circuits is in the area of refractory-metal deposition. For example, tungsten's low electrical resistivity (5.3 μΩ-cm) and its refractory nature make it a desirable metal for use in integrated circuits.

CVD Tungsten

Tungsten is used both as a contact plug and as a first-level metal. The CVD tungsten film is known for its excellent step coverage. For contact or via holes with size < 0.8 μm and aspect ratios greater than two, it is difficult to use conventional Al sputtering for continuous coating inside the feature and maintain the electrical performance. The effective via resistance and electromigration resistance have been improved by the introduction of CVD tungsten. The CVD tungsten process has been a key technology enabling multilevel interconnection metallization.

Tungsten can be deposited by using WF_6 as the W source gas, since it is a liquid that boils at room temperature. WF_6 can be reduced by silicon, hydrogen, or silane. The basic chemistry for CVD-W is as follows:

$$WF_6 + 3H_2 \rightarrow W + 6HF \text{ (hydrogen reduction)}, \tag{24}$$

$$2WF_6 + 3Si \rightarrow 2W + 3SiF_4 \text{ (silicon reduction)}, \tag{25}$$

$$2WF_6 + 3SiH_4 \rightarrow 2W + 3SiF_4 + 6H_2 \text{ (silane reduction)}. \tag{26}$$

On a Si contact, the selective process starts from a silicon reduction process. This process provides a nucleation layer of W grown on Si but not on SiO_2. The hydrogen reduction process can deposit W rapidly on the nucleation layer, forming the plug. The hydrogen reduction process provides excellent conformal coverage of the topography. This process, however, does not have perfect selectivity, and the HF gas by-product of the reaction is responsible for the encroachment of the oxide, as well as for the rough surface of deposited W films.

The silane reduction process gives a high deposition rate and much smaller W grain size than that obtained with the hydrogen reduction process. In addition, the problems of encroachment and rough W surface are eliminated because there is no HF by-product. Usually, a silane reduction process is used as the first step in blanket W deposition to serve as a nucleation layer and to reduce junction damage. After the silane reduction process, hydrogen reduction is used to grow the blanket W layer.

CVD TiN

Titanium nitride (TiN) is widely used as a diffusion barrier-metal layer in metallization and has numerous applications: (1) a cladding layer in Al metallization to enhance interconnection wiring electromigration resistance, (2) a CVD-W adhesion layer over oxide and a barrier against the interaction of WF_6 with Al and Si, (3) local interconnection where Al metallization cannot bear the temperature, (4) a plate electrode for Ta_2O_5 capacitors, and (5) the node and plate electrodes for MIM (metal-insulator-metal) capacitors where the insulator is either an atomic-layer-deposited Al_2O_3 or HfO_2/Al_2O_3 laminate.

TiN can be deposited by sputtering from a compound target or by CVD. The CVD TiN can provide better step coverage than PVD methods in deep submicron technology. CVD TiN can be deposited,[13–15] using $TiC1_4$ with NH_3, H_2/N_2, or NH_3/H_2:

$$6TiCl_4 + 8NH_3 \rightarrow 6\ TiN + 24HCl + N_2, \tag{27}$$

$$2TiCl_4 + N_2 + 4H_2 \rightarrow 2TiN + 8HCl, \tag{28}$$

$$2TiCl_4 + 2NH_3 + H_2 \rightarrow 2\ TiN + 8HCl. \tag{29}$$

The deposition temperature is about 400°–700°C for NH_3 reduction and is above 700°C for the H_2/N_2 reaction. The higher the deposition temperature, the better the TiN film and the less Cl incorporated in the TiN (~5%).

12.5.3 Aluminum Metallization

Aluminum and its alloys are used extensively for metallization in integrated circuits. The Al film can be deposited by a PVD or CVD method. Because aluminum and its alloys have low resistivities (2.7 μΩ-cm for Al and up to 3.5 μΩ-cm for its alloys), these metals satisfy the low-resistance requirements. Aluminum also adheres well to silicon dioxide. However, the use of aluminum in integrated circuits with shallow junctions often creates problems such as spiking and electromigration. We consider the problems of aluminum metallization and their solutions in this section.

CHAPTER 14

Impurity Doping

Impurity doping is the introduction of controlled amounts of impurity dopants into semiconductor materials. The practical use of impurity doping is primarily to change the electrical properties of the semiconductors. *Diffusion* and *ion implantation* are the two key methods of impurity doping. Until the early 1970s, impurity doping was done mainly by diffusion at elevated temperatures, as shown in Fig. 1*a*. In this method the dopant atoms are placed on or near the surface of the wafer by deposition from the gas phase of the dopant or by using doped-oxide sources. The doping concentration decreases monotonically from the surface, and the profile of the dopant distribution is determined mainly by the temperature and diffusion time.

Since the early 1970s, many doping operations have been performed by ion implantation, as shown in Fig. 1*b*. In this process the dopant ions are implanted into the semiconductor by means of an ion beam. The doping concentration has a peak distribution inside the semiconductor and the profile of the dopant distribution is determined mainly by the ion mass and the implanted-ion energy. Both diffusion and ion implantation are used in fabricating discrete devices and integrated circuits because these processes generally complement each other.[1,2] For example, diffusion is used to form a deep junction (e.g., a twin well in CMOS), whereas ion implantation is used to form a shallow junction (e.g.. a source/drain junction of a MOSFET).

Specifically, we cover the following topics:

- The movement of impurity atoms in the crystal lattice under high temperature and high concentration-gradient conditions.
- Impurity profiles for constant diffusivity and concentration-dependent diffusivity.
- The impact of lateral diffusion and impurity redistribution on device characteristics.
- The process and advantages of ion implantation.
- Ion distributions in the crystal lattice and how to remove lattice damage caused by ion implantation.
- Implantation-related processes such as masking, high-energy implantation, and high-current implantation.

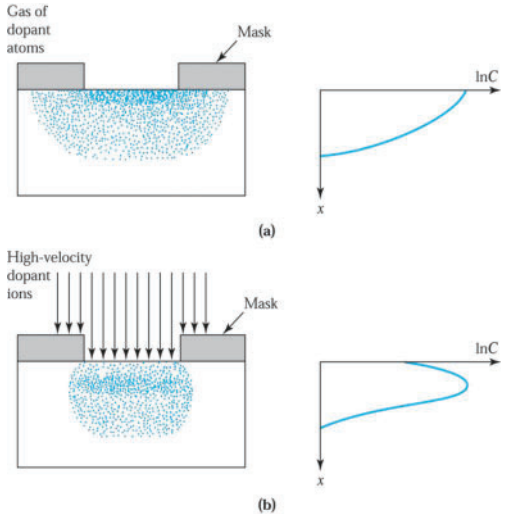


Fig. 1 Comparison of (*a*) diffusion and (*b*) ion-implantation techniques for selective introduction of dopants into the semiconductor substrate.

14.1 BASIC DIFFUSION PROCESS

Diffusion of impurities is typically done by placing semiconductor wafers in a carefully controlled high-temperature quartz-tube furnace and passing a gas mixture that contains the desired dopant through it. The temperature usually ranges between 800° and 1200°C for silicon and 600° and 1000°C for gallium arsenide. The number of dopant atoms that diffuse into the semiconductor is related to the partial pressure of the dopant impurity in the gas mixture.

For diffusion in silicon, boron is the most popular dopant for introducing a *p*-type impurity, whereas arsenic and phosphorus are used extensively as *n*-type dopants. These three elements are highly soluble in silicon, as they have solubilities above 5×10^{20} cm^{-3} in the diffusion temperature range. These dopants can be introduced in several ways, including solid sources (e.g., BN for boron, As_2O_3 for arsenic, and P_2O_5 for phosphorus), liquid sources (BBr_3 $AsCl_3$, and $POCl_3$), and gaseous sources (B_2H_6, AsH_3, and PH_3). However, liquid sources are most commonly used. A schematic diagram of the furnace and gas flow arrangement for a liquid source is shown in Fig. 2. This arrangement is similar to that used for thermal oxidation. An example of the chemical reaction for phosphorus diffusion using a liquid source is

$$4POCl_3 + 3O_2 \rightarrow 2P_2O_5 + 6Cl_2\uparrow. \quad (1)$$

The P_2O_5 forms a glass on a silicon wafer and is then reduced to phosphorus by silicon,

$$2P_2O_5 + 5Si \rightarrow 4P + 5SiO_2; \quad (2)$$

the phosphorus is released and diffuses into the silicon and Cl_2 is vented.

For diffusion in gallium arsenide, the high vapor pressure of arsenic requires special methods to prevent the loss of arsenic by decomposition or evaporation.[2] These methods include diffusion in sealed ampules with an overpressure of arsenic and diffusion in an open-tube furnace with a doped-oxide capping layer (e.g., silicon nitride). Most of the studies on *p*-type diffusion have been confined to the use of zinc in the forms of Zn-Ga-As alloys and $ZnAs_2$ for the sealed-ampule approach or ZnO-SiO_2 for the open-tube approach. The *n*-type dopants in gallium arsenide include selenium and tellurium.