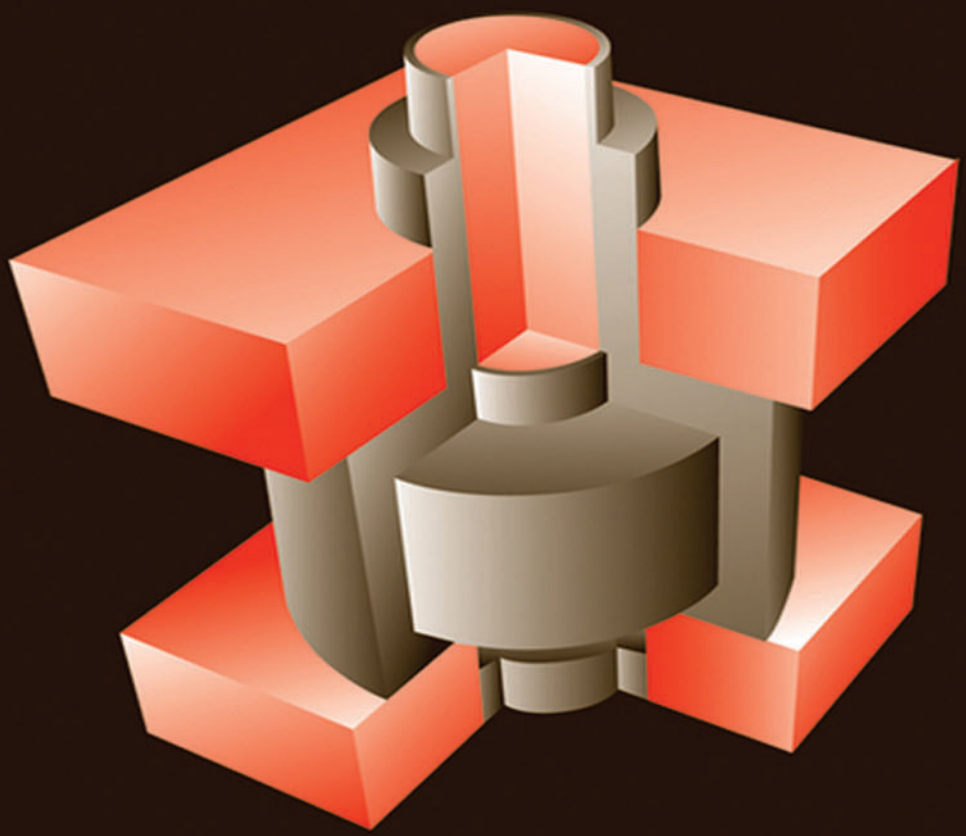
Semiconductor Devices

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.

Copyright © 1985, 2002, 2012 by John Wiley & Sons, Inc. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (508) 750-8400, fax (508) 750-4470. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc. 605 Third Avenue, New York, NY 10158-0012, (212) 850-6008, E-mail: PERMREQ@WILEY.COM. To order books or for customer service call 1-800-CALL-WILEY (225-5945).

Evaluation copies are provided to qualified academics and professionals for review purposes only, for use in their courses during the next academic year. These copies are licensed and may not be sold or transferred to a third party. Upon completion of the review period, please return the evaluation copy to Wiley. Return instructions and a free of charge return shipping label are available at www.wiley.com/go/returnlabel. Outside of the United States, please contact your local representative.

ISBN 978-0470-53794-7

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

0.2 SEMICONDUCTOR TECHNOLOGY

0.2.1 Key Semiconductor Technologies

Many important semiconductor technologies have been derived from processes invented centuries ago. For example, the lithography process was invented in 1798; in this first process, the pattern, or image, was transferred from a stone plate (lith- is Greek for 'stone').[26] In this section, we consider the milestones of technologies that were applied for the first time to semiconductor processing or were developed specifically for semiconductor-device fabrication.

Table 2 lists some key semiconductor technologies in chronological order. In 1918, Czochralski[27] developed a liquid-solid monocomponent growth technique. The Czochralski growth is the process used to grow most of the crystals from which silicon wafers are produced. Another growth technique, developed by Bridgman[28] in 1925, has been used extensively for growing gallium arsenide and related compound semiconductor crystals. Although the semiconductor properties of silicon have been widely studied since early 1940, the study of semiconductor compounds was neglected for a long time. In 1952, Welker[29] noted that gallium arsenide and its related III–V compounds were semiconductors. He was able to predict their characteristics and prove them experimentally. The technology and devices of these compounds have since been actively studied.

The diffusion of impurity atoms in semiconductors is important for device processing. Basic diffusion theory was considered by Fick[30] in 1855. The idea of using diffusion techniques to alter the type of conductivity in silicon was disclosed in a patent in 1952 by Pfann.[31] The lithography process was first applied to semiconductor-device fabrication by Andrus in 1957.[32] He used photosensitive etch-resistant polymers (photoresist) for pattern transfer. Lithography is a key technology for the semiconductor industry. The continued growth of the industry has been the direct result of improved lithographic technology. Lithography is also a significant economic factor, currently representing over 35% of the integrated-circuit manufacturing cost.

The oxide masking method was developed by Frosch and Derick[33] in 1957. They found that an oxide layer can prevent most impurity atoms from diffusing through it. In the same year, the epitaxial growth process based on chemical vapor deposition technique was developed by Sheftal et al.[34] Epitaxy (from the Greek epi, meaning on, and taxis, meaning arrangement) is a technique of crystal growth to form a thin layer of semiconductor materials on a crystal surface that has a lattice structure identical to that of the crystal. This method is important in improving device performance and creating novel device structures.

In 1958, Shockley[35] proposed a method of using ion implantation to dope semiconductors. Ion implantation has the capability of precisely controlling the number of implanted dopant atoms. Diffusion and ion implantation can complement each other for impurity doping. For example, diffusion can be used for high-temperature, deep-junction processes, whereas ion implantation can be used for lower-temperature, shallow-junction processes.

TABLE 2 KEY SEMICONDUCTOR TECHNOLOGIES

Year	Technology[a]	Author(s)/Inventor(s)	Ref.
1918	Czochralski crystal growth	Czochralski	27
1925	Bridgman crystal growth	Bridgman	28
1952	III-V compounds	Welker	29
1952	Diffusion	Pfann	31
1957	Lithographic photoresist	Andrus	32
1957	Oxide masking	Frosch and Derick	33
1957	Epitaxial CVD growth	Sheftal, Kokorish, and Krasilov	34
1958	Ion implantation	Shockley	35
1959	Hybrid integrated circuit	Kilby	36

1959	Monolithic integrated circuit	Noyce	37
1960	Planar process	Hoerni	38
1963	CMOS	Wanlass and Sah	39
1967	DRAM	Dennard	40
1969	Polysilicon self-aligned gate	Kerwin, Klein, and Sarace	41
1969	MOCVD	Manasevit and Simpson	42
1971	Dry etching	Irving, Lemons, and Bobos	43
1971	Molecular beam epitaxy	Cho	44
1971	Microprocessor (4004)	Hoff et al.	45
1981	Atomic layer deposition	Suntola	46
1982	Trench isolation	Rung, Momose, and Nagakubo	47
1989	Chemical mechanical polishing	Davari et al.	48
1993	Copper interconnect	Paraszczak et al.	49
2001	3D integration	Banerjee, et al.	50
2003	Immersion lithography	Owa, Nagasaka	51

[a]CVD, chemical vapor deposition; CMOS, complementary metal-oxide-semiconductor field-effect transistor; DRAM, dynamic random access memory; MOCVD, metalorganic CVD.

In 1959, a rudimentary integrated circuit (IC) was made by Kilby.[36] It contained one bipolar transistor, three resistors, and one capacitor, all made in germanium and connected by wire bonding—a hybrid circuit. Also in 1959, Noyce[37] created the monolithic IC by fabricating all devices in a single semiconductor substrate (monolith means 'single stone') and connecting the devices by aluminum metallization. Figure 6 shows the first monolithic IC of a flip-flop circuit containing six devices. The aluminum interconnection lines were obtained by etching evaporated aluminum layer over the entire oxide surface using the lithographic technique. These inventions laid the foundation for the rapid growth of the microelectronics industry.

The "planar" process was developed by Hoerni[38] in 1960. In this process, an oxide layer is formed on a semiconductor surface. With the help of a lithographic process, portions of the oxide can be removed and windows cut in the oxide. Impurity atoms will diffuse only through the exposed semiconductor surface, and *p–n* junctions will form in the oxide window areas.

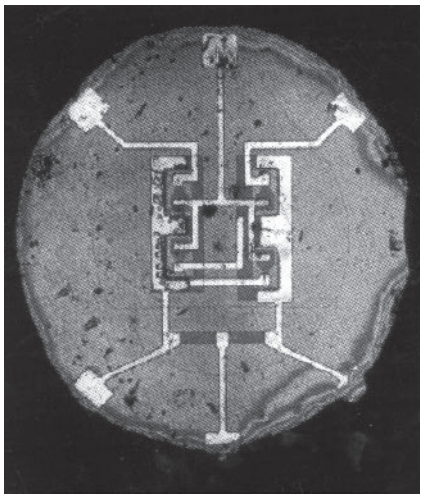


Fig. 6 The first monolithic integrated circuit.[37] (Photograph courtesy of Dr. G. Moore.)

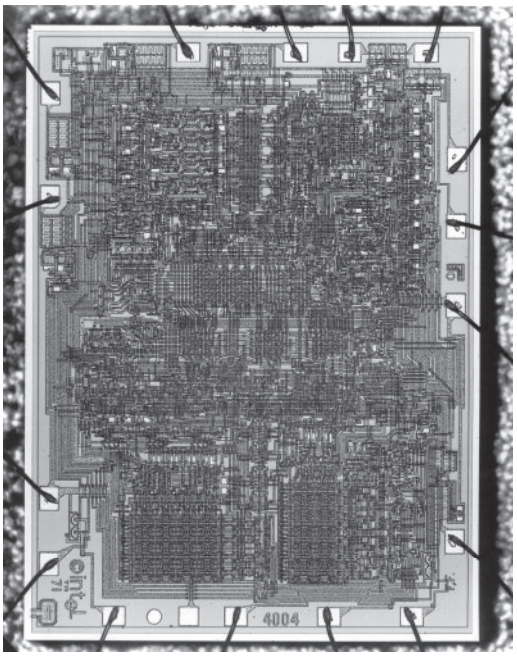


Fig. 7 The first microprocessor.[45] (Photograph courtesy of Intel Corp.)

As the complexity of the IC increased, we have moved from *NMOS* (*n*-channel MOSFET) to *CMOS* (complementary MOSFET) technology, which employs both NMOS and *PMOS* (*p*-channel MOSFET) to form the logic elements. The CMOS concept was proposed by Wanlass and Sah [39] in 1963. The advantage of CMOS technology is that logic elements draw significant current only during the transition from one state to another (e.g., from 0 to 1) and draw very little current between transitions, allowing power consumption to be minimized. CMOS technology is currently the dominant technology for advanced ICs.

In 1967, an important two-element circuit, the dynamic random-access memory (DRAM), was invented by Dennard.[40] The memory cell contains one MOSFET and one charge-storage capacitor. The MOSFET serves as a switch to charge or discharge the capacitor. Although DRAM is volatile and consumes relatively high power, we expect that DRAM will continue to be used in most electronic systems as an important working memory where information is held temporarily before being filed for long-term storage (e.g., in NVSM).

To improve the device performance, the polysilicon self-aligned gate process was proposed by Kerwin et al.[41] in 1969. This process not only improved device reliability but also reduced parasitic capacitances. Also in 1969, the metalorganic chemical vapor deposition (MOCVD) method was developed by Manasevit and Simpson.[42] This is a very important epitaxial growth technique for compound semiconductors such as GaAs.

As the device dimensions were reduced, a dry etching technique was developed to replace wet chemical etching for high-fidelity pattern transfer. This technique was initiated by Irving et al.[43] in 1971 using a $CF_4 - O_2$ gas mixture to etch silicon wafers. Another important technique developed in the same year by Cho is molecular beam epitaxy. [44] This technique has the advantage of near-perfect vertical control of composition and doping down to atomic dimensions. It is responsible for the creation of numerous photonic devices and quantum-effect devices.

In 1971, the first microprocessor was made by Hoff et al.[45] They put the entire central processing unit (CPU) of a simple computer on one chip. This was a four-bit microprocessor (Intel 4004), shown in Fig. 7, with a chip size of 3 mm × 4 mm, and it contained 2300 MOSFETs and operated at 0.1 MIPS (million instructions per second). It was fabricated by a *p*-channel, polysilicon gate process using an 8 μm design rule. This microprocessor performed as well as those in $300,000 IBM computers of the early 1960s—each of which needed a CPU the size of a large desk. This was a major breakthrough for the semiconductor industry. Currently, microprocessors constitute the largest segment of the industry.

Since early 1980, many new technologies have been developed to meet the requirements of ever-shrinking minimum feature lengths. An important technique, atomic layer deposition (ALD), was developed for nanoscaled dielectric film deposition by Suntola in 1981.[46] This deposition technique involves exposing the chemical precursors to the growth surface in a sequential, one-at-a-time manner. The film thickness can be reliably controlled down to atomic dimensions.

The trench isolation technology was introduced by Rung et al.[47] in 1982 to isolate CMOS devices and has ultimately replaced all other isolation methods. In 1989, Davari et al.[48] developed the chemical-mechanical polishing method for global planarization of the interlayer dielectrics. This is a key process in multilevel metallization.

At submicron dimensions, a widely known failure mechanism is electromigration, which is the transport of metal ions through a conductor due to the passage of an electrical current. Although aluminum has been used since the early 1960s as the interconnect material, it suffers from electromigration at high electrical current. The copper interconnect was introduced in 1993 by Paraszczak et al.[49] to replace aluminum for minimum feature lengths below 100 nm.

The increased component density and improved fabrication technology have helped realize the system-on-a-chip (SOC), which is an IC chip containing a complete electronic system. The system-on-a-chip was integrated into a three-dimensional (3-D) system with improved performance by Banerjee in 2001.[50]

In order to extend the optical photolithography to the nanoscale regime, Owa et al. in 2003[51] developed immersion lithography through the addition of water between the exposure lens and the wafer surface. Immersion lithography increases the resolution by a factor equal to the refractive index of the liquid, and the minimum feature size can be made below 45 nm. In Part III of this book, we consider all the technologies listed in Table 2.

0.2.2 Technology Trends

Since the beginning of the microelectronics era, the smallest line width or the minimum feature length of an integrated circuit has been reduced at a rate of about 13% per year.[52] At that rate, the minimum feature length will shrink to about 10 nm in the year 2020. Figure 8 shows the minimum feature length versus year of first production from 1978 to 2010 and projected to 2020. In 2002 we entered the nanoelectronics era by reducing the minimum feature length below 100 nm.

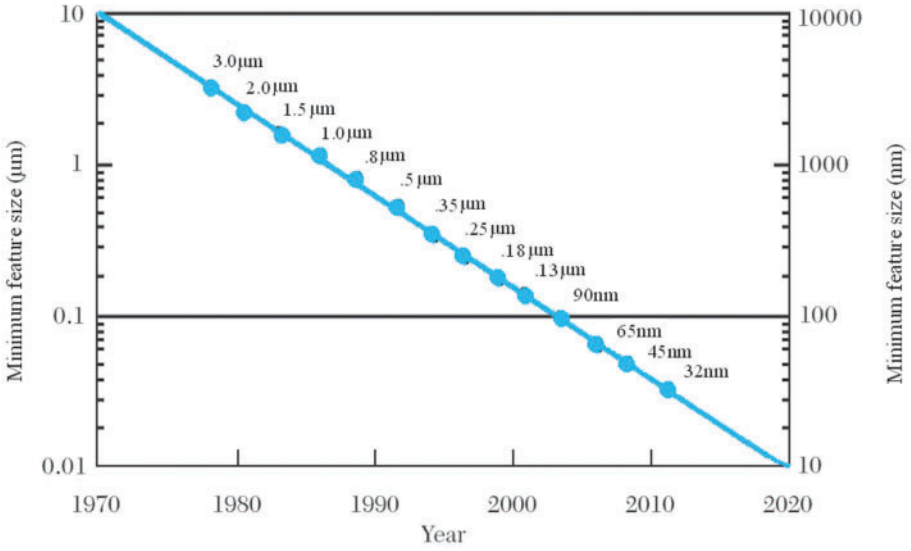


Fig. 8 **Exponential decrease of the minimum feature length versus time.**[52]

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.

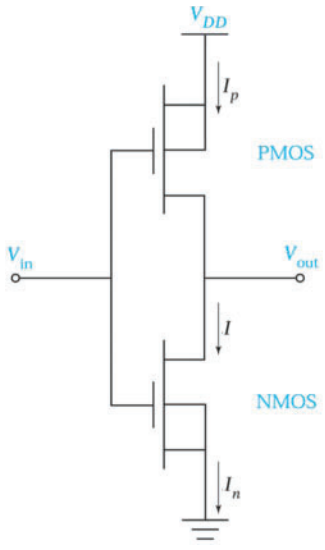


Fig. 12 **The CMOS inverter.**

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

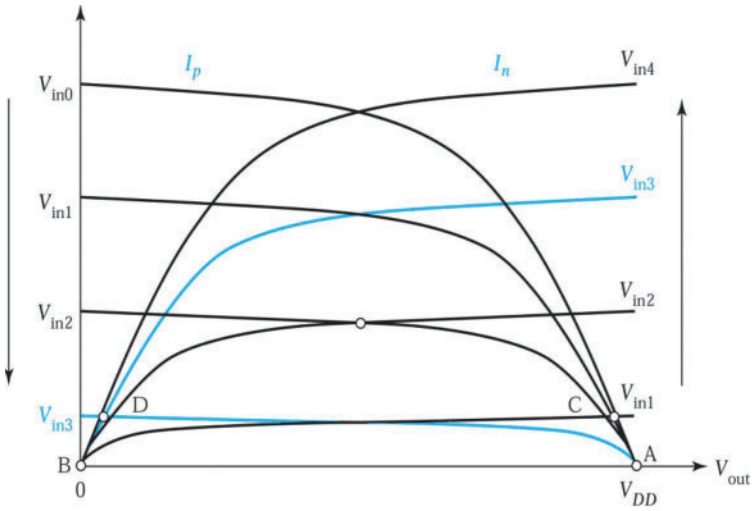


Fig. 13 **I_p and I_n as functions of V_{out}. The intercepts of I_p and I_n (circled) represent the steady-state operation points of the CMOS inverter.[4] The curves are labeled by the input voltages: $0 = V_{in0} < V_{in1} < V_{in2} < V_{in3} < V_{in4} = V_{DD}$.**

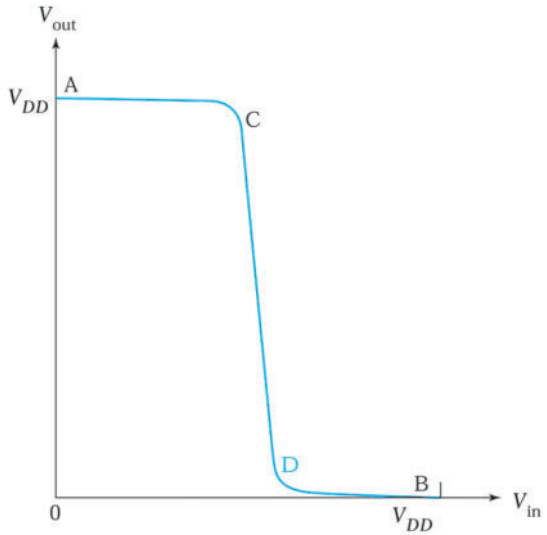


Fig. 14 **Transfer curve of a CMOS inverter.[4] Points labeled A, B, C, and D correspond to the points labeled in Fig. 13.**

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.

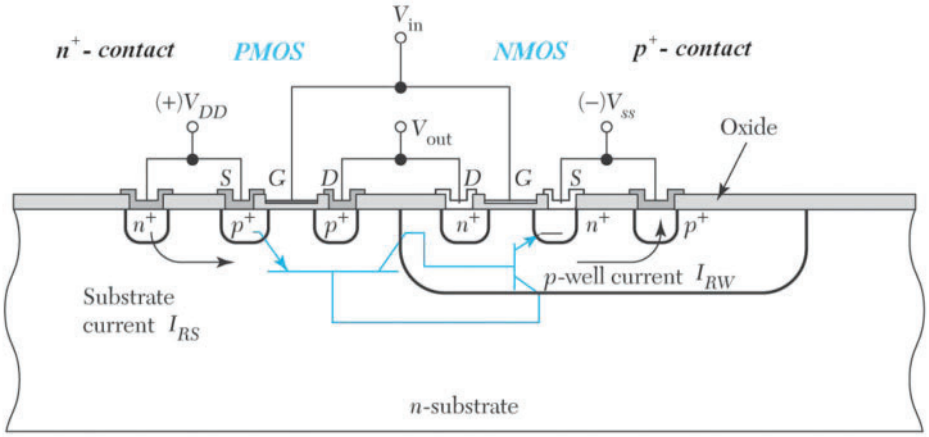


Fig. 15 **Cross section of a CMOS inverter fabricated with *p*-well technology.**

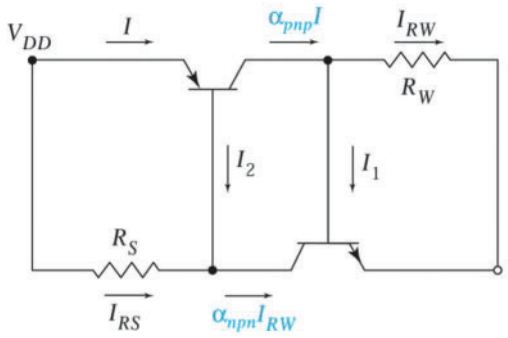


Fig. 16 **Equivalent circuit of the *p*-well structure shown in Fig. 15.**

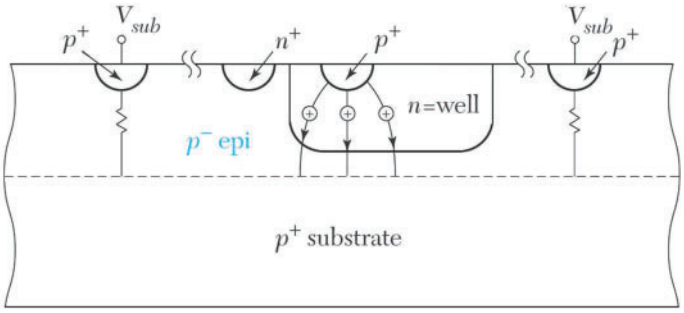


Fig. 17 **Prevention of latch-up with a heavily doped substrate.**[7]

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.

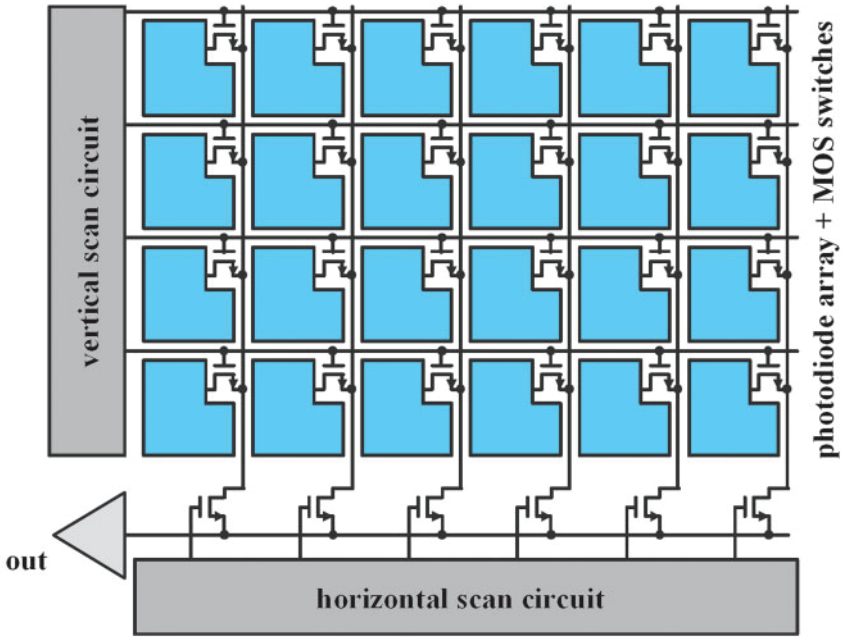


Fig. 18 **Architecture of a two-dimensional CMOS image sensor.**

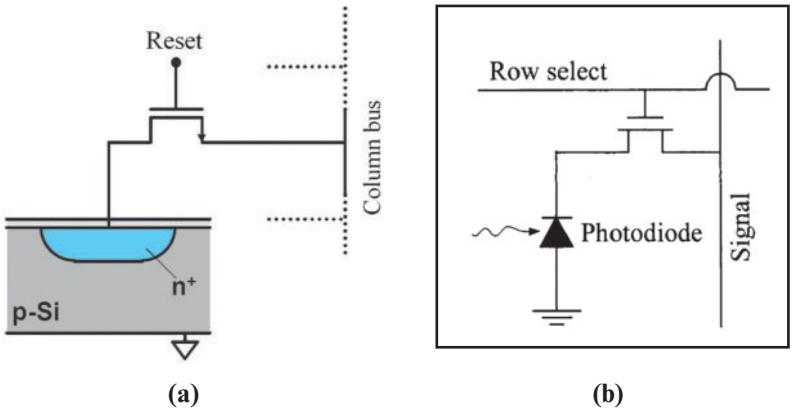


Fig. 19 **(*a*) Passive CMOS pixel based on a single in-pixel transistor. (*b*) PPS (passive pixel sensor) CMOS image sensors.**

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.

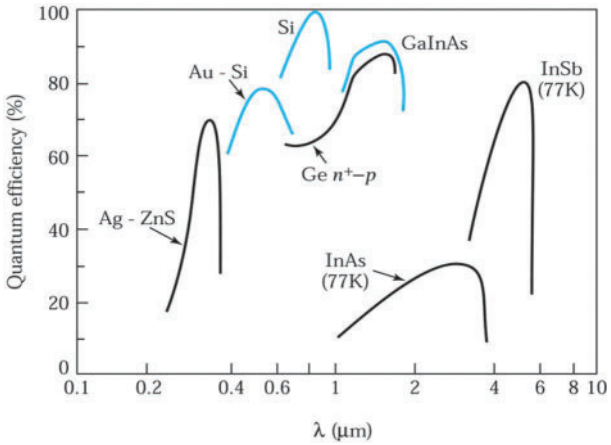


Fig. 2 Quantum efficiency versus wavelength for various photodetectors.[1,2]

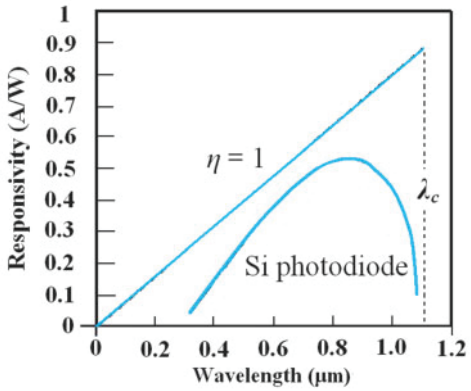


Fig. 3 **Responsivity vs. wavelength for an ideal photodiode with $\eta = 1$ and for a typical commercial Si photodiode.**

Figure 2 shows typical plots of quantum efficiency versus wavelength for some high-speed photodiodes.[1,2] Note that in the ultraviolet and visible region, metal-semiconductor photodiodes (discussed in Section 10.1.4) show good quantum efficiencies. In the near-infrared region, silicon photodiodes (with an antireflection coating) can reach 100% quantum efficiency near the 0.8- to 0.9-μm region. In the 1.0- to 1.6-μm region, germanium photodiodes and Group III-V photodiodes (e.g., GaInAs) have shown high quantum efficiencies. For even longer wavelengths, photodiodes are cooled (e.g., to 77 K) for high-efficiency operation.

Responsivity

The responsivity $\mathscr{R}$ of a photodiode is defined as the generated photocurrent (I_p) per incident optical power (P_{opt}). $\mathscr{R}$ is also called the *spectral responsivity* or *radiant sensivity*:

$$\mathscr{R} = I_p / P_{opt} \tag{10}$$

From the definition of quantum efficiency, we have

$$\mathscr{R} = I_p / P_{opt} = \eta q / h\nu = \eta q \lambda / hc \tag{11}$$

If a photodiode has an ideal quantum efficiency of 100%, then $\mathscr{R}$ should be linearly proportional to the wavelength. In practice, the relationship of $\mathscr{R}$ and λ is shown in Fig. 3. The quantum efficiency limits the responsivity below the ideal photodiode.

Response Speed

The response speed is limited by three factors: (1) diffusion of carriers, (2) drift time in the depletion region, and (3) capacitance of the depletion region. Carriers generated outside the depletion region must diffuse to the junction, resulting in considerable time delay. To minimize the diffusion effect, the junction should be formed very close to the surface. The greatest amount of light will be absorbed when the depletion region is wide. However, the depletion layer must not be too wide or transit time effects will limit the frequency response. It also should not be too thin, or excessive capacitance C will result in a large RC time constant, where R is the load resistance. The optimal compromise is the width at which the depletion layer transit time is approximately one half the modulation period. For example, for a modulation frequency of 2 GHz, the optimal depletion-layer thickness in silicon (with a saturation velocity of 10^7 cm/s) is about 25 μm.

10.1.3 *p-i-n* Photodiode

As described above, the *p-n* junction photodiode has two major drawbacks. First, the junction capacitance is not sufficiently small, due to the small depletion layer width. For example, the depletion layer width is below 1 μm for a p^+-*n* silicon junction as in Ex. 2 in Chapter 3. It contributes a large *RC* time constant so that the photodiode cannot operate at high modulation frequencies. In addition, its depletion layer is not sufficiently wide to make the penetration depth greater than the depletion layer width at long wavelengths. The penetration depth is about 33 μm, for example, at the wavelength 900 nm shown in Fig. 5, Chapter 9. Most incident photons are absorbed outside the depletion region where there is no field to separate the EHPs.

The *p-i-n* (*p-intrinsic-n*) photodiode is one of the most common photodetectors, because the depletion layer thickness (the intrinsic layer) can be tailored to optimize the quantum efficiency and frequency response. The *i*-layer thickness is typically 5~50 μm depending on the particular application. The intrinsic *i*-layer in a *p-i-n* photodiode is completely depleted. The junction capacitance is sufficiently small due to the large depletion layer width to make the *p-i-n* photodiode operate at high modulation frequencies. Its depletion layer is also wide enough to have a large absorption in the depletion layer at long wavelengths.

Figure 4*a* shows a cross section of a *p-i-n* photodiode that has an antireflection coating to increase quantum efficiency. The surface reflection of the incident light from air ($n = 1$) into semiconductor silicon ($n = 3.5$) is about 0.31, from Eq. 22 in Chapter 9. This means that 31% of incident light is reflected and is not available for conversion to electrical energy. Covering the surface with an antireflection coating with a refractive index $n = (n_{Si})^{1/2}$ minimizes the total reflection. Si_3N_4 with $n = 1.9$ is a good choice. Figures 4*b* shows the energy band diagram of the *p-i-n* diode under reverse bias condition. The conduction band decreases linearly with distance and the electric field is uniform in the *i*-layer.

The optical absorption is shown in Fig. 4*c*. The *p-i-n* structure is designed so that almost complete optical absorption occurs over the *i*-layer. The EHPs produced in the depletion region or within a diffusion length of it from light absorption will eventually be separated by the electric field and a current flows in the external circuit as carriers drift across the depletion layer.

Generally, the response time is limited by the drift time of the slowest photogenerated carriers, holes, across the width of *i*-layer. A narrower *i*-layer improves the response time but decreases the quantity of absorbed photons and hence reduces the responsivity. To increase the response speed, i.e., to reduce the drift time, we have to increase the reverse bias. There is therefore a trade-off between speed and responsivity.

In practice, the *i*-layer will have a slight background doping, The structure is more like p^+-π-n^+ or p^+-ν-n^+ mentioned in Fig. 27 in Chapter 3. The field is not uniform across the *i*-layer. As an approximation, we can still consider it as a *p-i-n* structure.

▶ EXAMPLE 2

On reaching the surface of the semiconductor, the incident optical power P_0 will have its level reduced to $P_0(1 - R)$ on entering the material, where R is the reflection coefficient. On passing though the semiconductor the light will be absorbed, and so at any depth x the amount of residual optical power $P(x)$ is given by $P(x) = P_0(1 - R)\exp(-\alpha x)$. For $\alpha = 10^4$ cm^{-1} and $R = 0.1$, calculate the depth at which half the incident optical power has been absorbed in a material.

SOLUTION

$$x = \frac{-1}{\alpha}\ln\left[\frac{P(x)}{P_0(1-R)}\right] = -10^{-4}\cdot\ln\left(\frac{1}{2\times 0.9}\right)\ \text{cm}$$

$$= 0.59\ \mu\text{m}\ .$$

◀

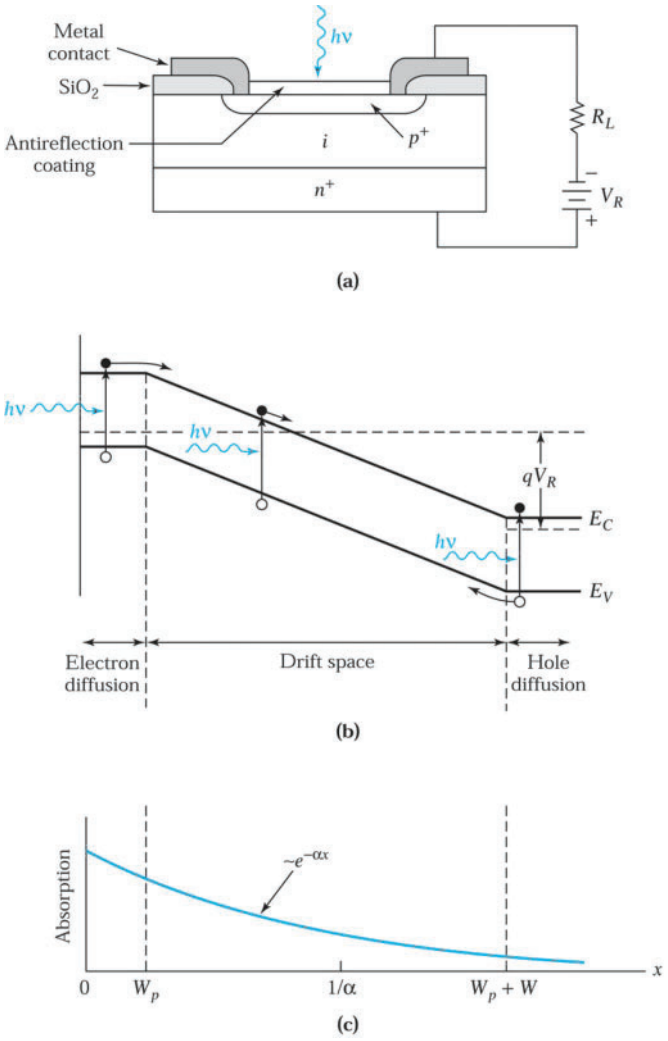


Fig. 4 Operation of a *p-i-n* photodiode. (*a*) Cross-sectional view of a *p-i-n* photodiode. (*b*) Energy band diagram under reverse bias. (*c*) Carrier absorption characteristics.

▶ EXAMPLE 3

The diameter of the optical receiving area of a Si *p-i-n* photodiode is 0.06 cm. It is illuminated with an incident optical intensity of 0.2 mW/cm^2 at wavelength 800 nm to generate a photocurrent of 3×10^{-4} mA. What are the responsivity and quantum efficiency of the *p-i-n* photodiode at 800 nm?

SOLUTION

The incident optical intensity is 0.2 mW/cm^2 and the diameter of optical receiving area is 0.06 cm. Thus, the incident power is

$$P_{opt} = \pi\,(0.03\text{cm})^2 \times 0.2\ \text{mW/cm}^2 = 5.6 \times 10^{-4}\ \text{mW}$$

The responsivity is

$$\mathscr{R} = I_p / P_{opt} = 3 \times 10^{-4}\ \text{mA}/5.6 \times 10^{-4}\ \text{mW} = 0.54\ \text{A/W}$$

The quantum efficiency is

$$\eta = \mathscr{R}(hc/q\lambda) = 0.54 \text{ A/W } (6.62 \times 10^{-34} \text{ J-s})(3 \times 10^{8} \text{ m/s})/(1.6 \times 10^{-19} \text{ C})(80 \times 10^{-9} \text{m}) = 0.84 = 84\% .$$ ◀

10.1.4 Metal-Semiconductor Photodiode

The construction of a high-speed metal-semiconductor (M-S) photodiode is shown in Fig. 5. To avoid large reflection and absorption losses when the diode is illuminated through the metal contact, the metal film must be very thin (~10 nm) and an antireflection coating must be used. Metal-semiconductor (M-S) photodiodes are particularly useful in the ultraviolet- and visible-light regions. In these regions the absorption coefficients, α, in most common semiconductors are very high, of the order of 10^5 cm^{-1} or more, which corresponds to an effective absorption length $1/\alpha$ of 0.1 μm or less. It is possible to choose a metal and an antireflection coating so that a large fraction of the incident radiation will be absorbed near the surface of the semiconductor. As an example, for a gold-silicon photodetector having 10 nm gold and 50 nm zinc sulfide as the antireflection coating, more than 95% of the incident light with $\lambda = 0.6328$ μm (helium-neon laser wavelength, red light) will be transmitted into the silicon substrate.

The M-S photodiode can be operated in two modes, depending on the photon energy. For $h\nu > E_g$ (Fig. 6*a*), the radiation produces EHPs in the semiconductor, and the general characteristics of the M-S photodiode are very similar to those of a *p-i-n* photodiode. For smaller photon energy (longer wavelength) $q\phi_B < h\nu < E_g$ (Fig. 6*b*), the photoexcited electrons in the metal can surmount the barrier and be collected by the semiconductor. This process is called *internal photoemission* and has been used extensively to determine the Schottky-barrier height and to study the hot electron transport in metal films.

When a Schottky-barrier diode is scanned with light of variable wavelength, Fig. 6*c* shows that the quantum efficiency has a threshold of $q\phi_B$ that increases with the photon energy. When the photon energy reaches the energy-gap value, the quantum efficiency jumps to a much higher value. In practical applications, however, the internal photoemission has typical quantum efficiencies of only less than 1%.

For detectors with internal photoemission, it is more efficient to direct the incoming light through the substrate. Since the barrier height is always smaller than the energy gap, light with $q\phi_B < h\nu < E_g$ is not absorbed in the semiconductor, and the intensity is not reduced at the metal/semiconductor interface. The metal layer, in this case, can be thicker for easier thickness control and to minimize series resistance. For Si devices, options are available using silicides in place of the metal. A silicide usually has a more reproducible interface since it is formed by reacting metal with Si so that the new interface is never exposed. Common silicides used for this purpose are PtSi, Pd_2Si, and IrSi. Another advantage of a Schottky-barrier diode is that it does not require high-temperature processing for diffusion or implantation annealing.

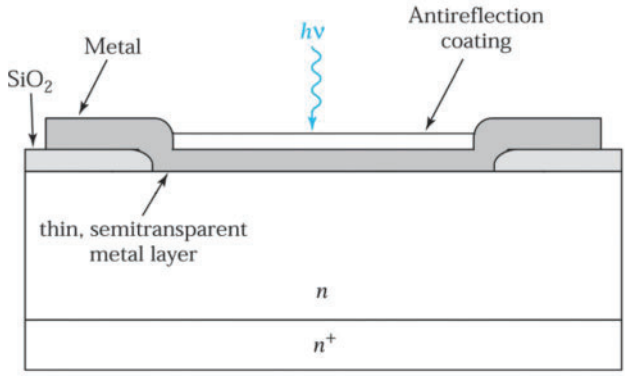


Fig. 5 **Metal-semiconductor photodiode.**

CHAPTER 12

Film Formation

- 12.1 THERMAL OXIDATION
- 12.2 CHEMICAL VAPOR DEPOSITION OF DIELECTRICS
- 12.3 CHEMICAL VAPOR DEPOSITION OF POLYSILICON
- 12.4 ATOMIC LAYER DEPOSITION
- 12.5 METALLIZATION
- SUMMARY

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

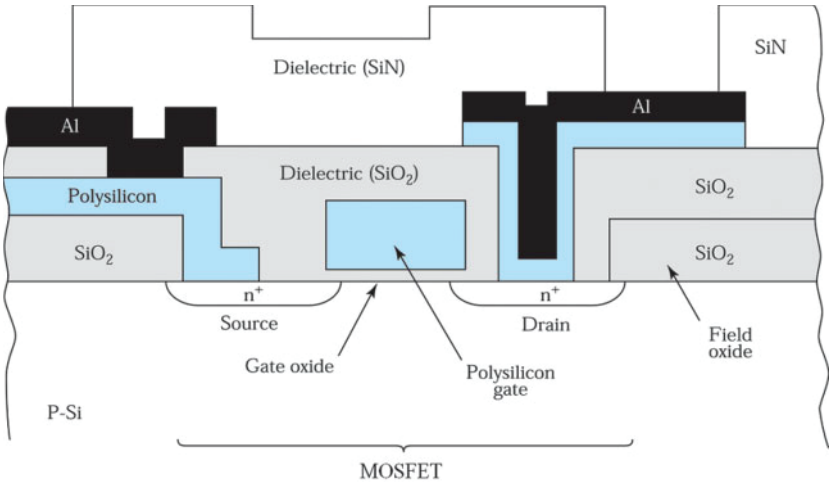


Fig. 1 Schematic cross section of a metal-oxide-semiconductor field-effect transistor (MOSFET).

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.

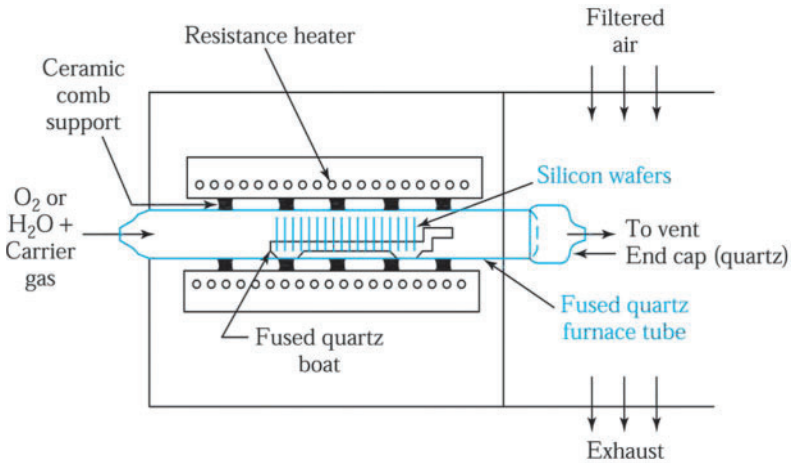


Fig. 2 Schematic cross section of a resistance-heated oxidation furnace.

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. ◀

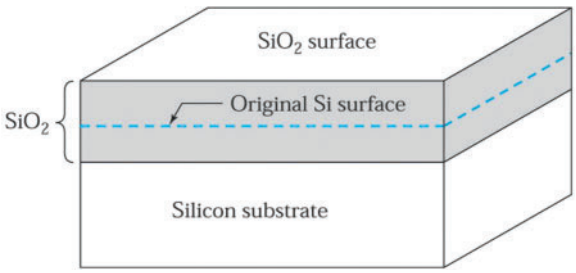


Fig. 3 **Growth of silicon dioxide by thermal oxidation.**

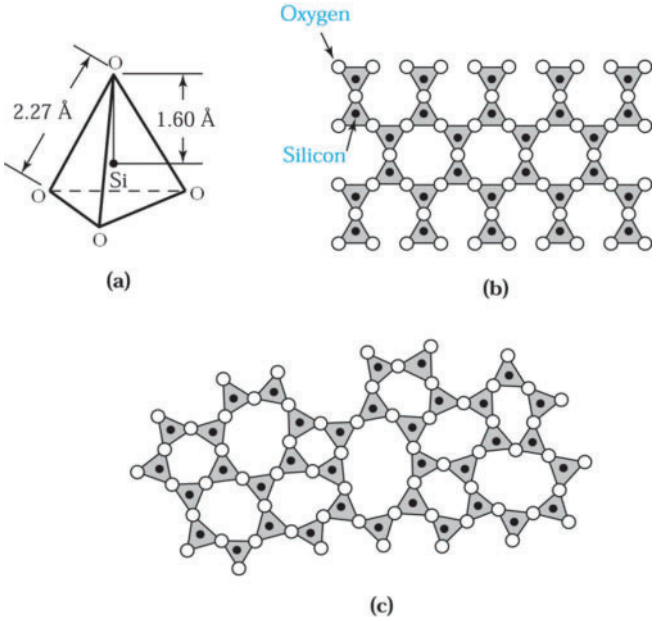


Fig. 4 (*a*) Basic structural unit of silicon dioxide. (*b*) Two-dimensional representation of a quartz crystal lattice. (*c*) Two-dimensional representation of the amorphous structure of silicon dioxide.[1]

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.

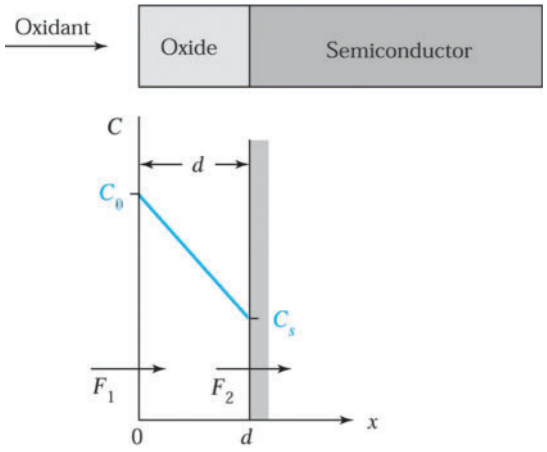


Fig. 5 **Basic model for the thermal oxidation of silicon.**[2]

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.

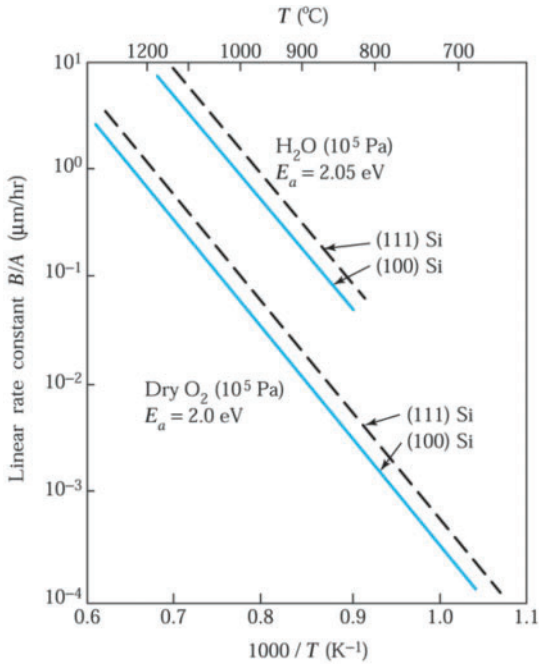


Fig. 6 **Linear rate constant versus temperature.**[2]

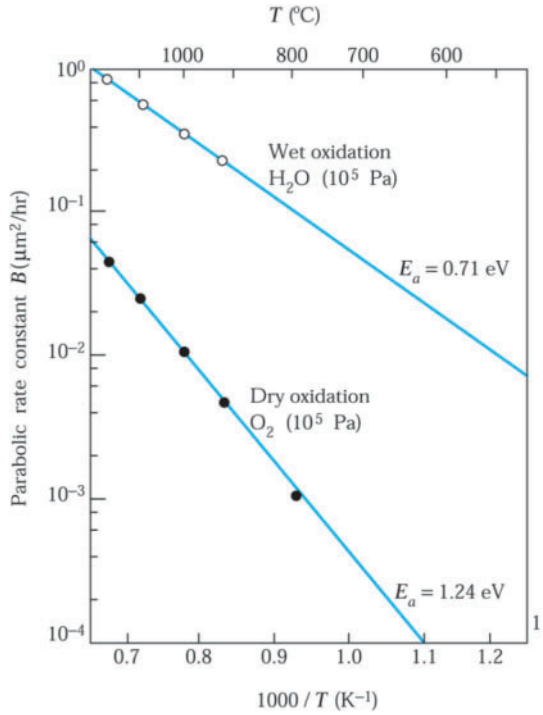


Fig. 7 **Parabolic rate constant versus temperature.**[2]

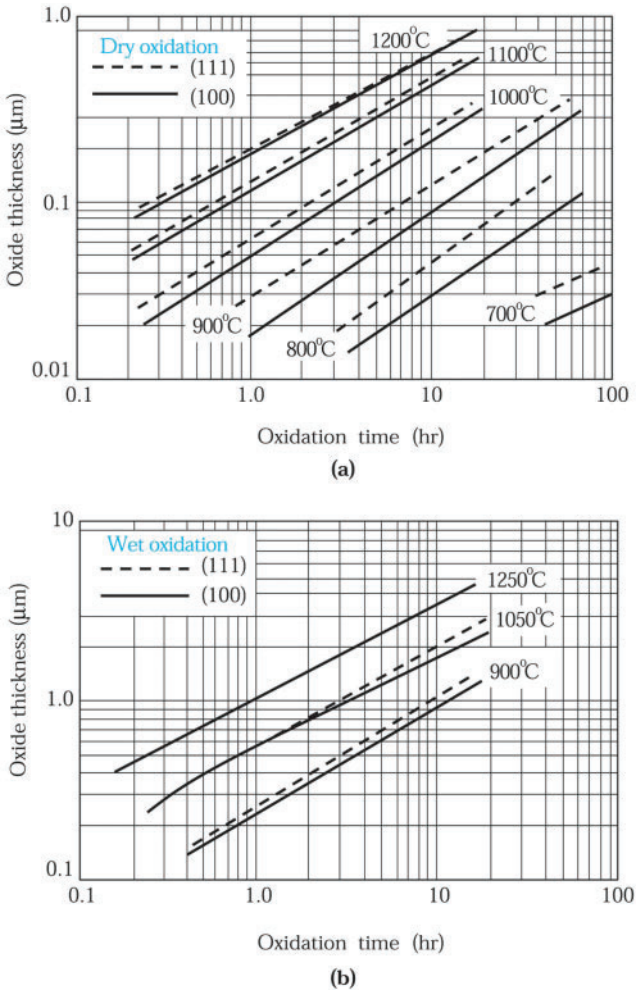


Fig. 8 Experimental results for silicon dioxide thickness as a function of reaction time and temperature for two substrate orientations. (*a*) Growth in dry oxygen. (*b*) Growth in steam.[3]

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

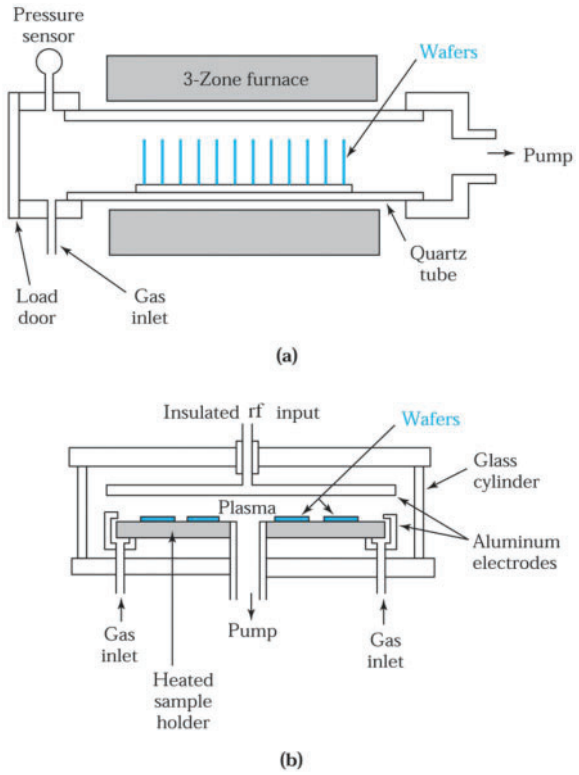


Fig. 9 Schematic diagrams of chemical-vapor deposition reactors. (*a*) Hot-wall LPCVD reactor. (*b*) Parallel-plate rf (radio frequency) plasma deposition reactor.[4]

rates. In a hot-wall LPCVD reactor as shown in Fig. 9*a*, the quartz tube is heated by a three-zone furnace, and gas is introduced at one end and pumped out at the opposite end. The semiconductor wafers are held vertically in a slotted quartz boat.[4] The quartz tube wall is hot because it is adjacent to the furnace, in contrast to a cold-wall reactor such as the horizontal epitaxial reactor, which uses radio frequency (rf) heating. The choice of a hot-wall or cold-wall reactor depends on whether the reaction is exothermic or endothermic. For the exothermic reaction, the deposition rate is lower with increasing temperature. These processes require a hot-wall reactor. However, in a cold-wall reactor, deposition would occur on the cooler reactor walls. Consequently, for the endothermic reaction, a cold-wall reactor is used. The deposition rate is higher on the substrates with higher temperatures.

PECVD is an energy-enhanced CVD method in which plasma energy is added to the thermal energy of a conventional CVD system. The parallel-plate, radial-flow PECVD reactor shown in Fig. 9*b* consists of a cylindrical glass or aluminum chamber sealed with aluminum endplates. Inside are two parallel aluminum electrodes. An rf voltage is applied to the upper electrode, whereas the lower electrode is grounded. The rf voltage causes a plasma discharge between the electrodes. Wafers are placed on the lower electrode, which is heated to between 100° and 400°C by resistance heaters. The reaction gases flow through the discharge from inlets located along the circumference of the lower electrode. The main advantage of this reactor is its low deposition temperature. However, its capacity is limited, especially for large-diameter wafers, and the wafers may become contaminated if loosely adhering deposits fall on them.

The substrate surface not only receives active precursors but is subject to the bombardment of charged species. The short-lived active species react and deposit on the surface, while the thermal energy and ion bombardment continue to modify the deposited materials. The plasma-enhanced deposited films tend to be of smaller grain size or even amorphous, and contain amounts of impurities such as hydrogen, carbon or halide atoms.

The combination of low temperature, self-cleaning capability, and versatile film tunability has assured the importance of PECVD in the semiconductor industry. To minimize deposits on the reactor surfaces, limiting the plasma area is beneficial. The standard parallel-plate configuration provides an efficient design to focus the deposition on the wafer. At the same time, the reactor's plasma capability also provides the potential for in-situ plasma cleaning by introducing etchant cleaning gases such as C_2F_6 or NF_3 to remove silicon dioxide and silicon nitride deposition from chamber surfaces. One limitation of plasma deposition involves the potential charge imbedded in the film.

To overcome charge damage and still maintain the advantages of a low-temperature process, remote plasma instead of in-situ plasma is used. Reactants are plasma dissociated or activated remotely, then introduced onto the substrate surface along with second reactants to complete the reaction. But one has to consider the short lifetime of the activated species and how to distribute them over the large substrate surface. There is one closely related successful example, TEOS/O_3. Fortunately, the O_3 is stable enough and the concentration can be high enough to produce a reasonable silica deposition rate and provide good step coverage.

CVD Processes

Chemical vapor deposition (CVD) is a method of forming a thin solid film on a substrate by the reaction of vapor-phase chemicals that contain the required constituents. The CVD process can be generalized in a sequence of steps. (1) Reactants are introduced into the reactor; (2) Gas species are activated and dissociated by mixing, heating, plasma, or other means; (3) Reactive species are adsorbed on the substrate surface; (4) Adsorbed species undergo chemical reaction or react with other incoming species to form a solid film; (5) Reaction byproducts are desorbed from the substrate surface; (6) Reaction byproducts are removed from the reactor.

Although film growth is primarily accomplished at step 4, the overall growth rate is controlled by steps 1-6 in series. The slowest step determines the final growth rate. As in any typical chemical kinetics, the determining factors are the concentrations of surface species, wafer temperature, and incoming charged species and their energies. Chemical vapor deposition process parameters must be finely adjusted to meet all the film properties and production requirements.

12.2.2 Silicon Dioxide

CVD silicon dioxide does not replace thermally grown oxides because the best electrical properties are obtained with thermally grown films. CVD oxides are used instead to complement thermal oxides. A layer of undoped silicon dioxide is used to insulate multilevel metallization, to mask ion implantation and diffusion, and to increase the thickness of thermally grown field oxides. Phosphorus-doped silicon dioxide is used both as an insulator between metal layers and as a final passivation layer over devices. Oxides doped with phosphorus, arsenic, or boron are used occasionally as diffusion sources.

Deposition Methods

Silicon dioxide films can be deposited by several methods. For low-temperature deposition (300°-500°C), the films are formed by reacting silane (SiH_4), dopant, and oxygen. The chemical reactions for phosphorus-doped oxides are

$$SiH_4 + O_2 \xrightarrow{450^{\circ}C} SiO_2 + 2H_2, \tag{15}$$

$$4PH_3 + 5O_2 \xrightarrow{450^{\circ}C} 2P_2O_5 + 6H_2, \tag{16}$$

The deposition process can be performed either at atmospheric pressure CVD or at LPCVD (Fig. 9*a*). The low deposition temperature of the silane-oxygen reaction makes it a suitable process when films must be deposited over a layer of aluminum.

For intermediate-temperature deposition (500°-800°C), silicon dioxide can be formed by decomposing tetraethylorthosilicate, $Si(OC_2H_5)_4$, in an LPCVD reactor. The compound, abbreviated TEOS, is vaporized from a liquid source. The TEOS compound decomposes as follows:

$$Si(OC_2H_5)_4 \xrightarrow{700^{\circ}C} SiO_2 + \text{byproducts}, \tag{17}$$

forming both SiO_2 and a mixture of organic and organosilicon byproducts. Although the higher temperature required for the reaction prevents its use over aluminum, it is suitable for polysilicon gates requiring a uniform insulating layer with good step coverage. The good step coverage is a result of enhanced surface mobility at higher temperatures. The oxides can be doped by adding small amounts of the dopant hydrides (phosphines, arsine, or diborane), similar to the process in epitaxial growth.

The deposition rate as a function of temperature varies as $e^{-Ea/kT}$, where E_a is the activation energy. The E_a of the silane-oxygen reaction is quite low: about 0.6 eV for undoped oxides and almost zero for phosphorus-doped oxide. In contrast, E_a for the TEOS reaction is much higher: about 1.9 eV for undoped oxide and 1.4 eV when phosphorus doping compounds are present. The dependence of the deposition rate on TEOS partial pressure is proportional to $(1-e^{-P/P_0})$, where P is the TEOS partial pressure and P_0 is about 30 Pa. At low TEOS partial pressures, the deposition rate is determined by the rate of the surface reaction. At high partial pressures, the surface becomes nearly saturated with adsorbed TEOS and the deposition rate becomes essentially independent of TEOS pressure.[4]

Recently, atmospheric-pressure and low-temperature CVD processes using TEOS and ozone (O_3) have been proposed.[5] This CVD technology produces oxide films with high conformality and low viscosity at low deposition temperatures. Because of their porosity, TEOS/O_3 CVD oxides are often accompanied by plasma-assisted oxides to permit planarization in ULSI processing.

For high-temperature deposition (900°C), silicon dioxide is formed by reacting dichlorosilane, $SiCl_2H_2$, with nitrous oxide at reduced pressure:

$$SiCl_2H_2 + 2N_2O \xrightarrow{900^{\circ}C} SiO_2 + 2N_2 + 2HCl. \quad (18)$$

This deposition gives excellent film uniformity and is sometimes used to deposit insulating layers over polysilicon.

Properties of Silicon Dioxide

Deposition methods and properties of silicon dioxide films are listed[4] in Table 1. In general, there is a direct correlation between deposition temperature and film quality. At higher temperatures, deposited oxide films are structurally similar to silicon dioxide that has been thermally grown.

The lower densities occur in films deposited below 500°C. Heating deposited silicon dioxide at temperatures between 600° and 1000°C causes densification, during which the oxide thickness decreases whereas the density increases to 2.2 g/cm^3. The refractive index of silicon dioxide is 1.46 at a wavelength of 0.6328 μm. Oxides with lower indices are porous, such as the oxide from the silane-oxygen deposition, which has a refractive index of 1.44. The porous nature of the oxide also is responsible for the lower dielectric strength and hence a higher leakage current in the oxide film. The etch rates of oxides in a hydrofluoric acid solution depend on deposition temperature, annealing history, and dopant concentration. Usually higher-quality oxides are etched at lower rates.

TABLE 1 PROPERTIES OF SiO_2 FILMS

Property	Thermally grown at 1000°C	$SiH_4 + O_2$ at 450°C	TEOS at 700°C	$SiCl_2H_2 + N_2O$ at 900°C
Composition	SiO_2	SiO_2 (H)	SiO_2	SiO_2 (Cl)
Density(g/cm^3)	2.2	2.1	2.2	2.2
Refractive index	1.46	1.44	1.46	1.46
Dielectric strength (10^6 V/cm)	>10	8	10	10
Etch rate (Å/min) (100:1 H_2O:HF)	30	60	30	30
Etch rate (Å /min) (buffered HF)	440	1200	450	450
Step coverage	—	Nonconformal	Conformal	Conformal

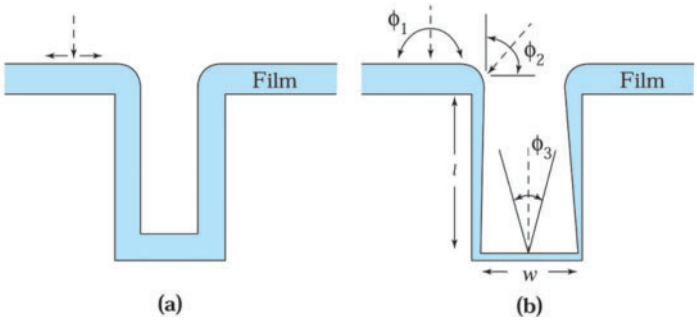


Fig. 10 Step coverage of deposited films. (*a*) Conformal step coverage. (*b*) Nonconformal step coverage.[4]

Step Coverage

Step coverage relates the surface topography of a deposited film to the various steps on the semiconductor substrate. Step coverage is one of the main advantages of the CVD method, especially compared with PVD. To get good step coverage, the inherent chemistries and operating conditions are critical. In the illustration of ideal, or conformal, step coverage shown in Fig. 10*a*, film thickness is uniform along all surfaces of the step. The uniformity of the film thickness, regardless of topography, is due to the rapid migration of reactants after adsorption on the step surfaces.[6]

Figure 10*b* shows an example of nonconformal step coverage, which results when the reactants adsorb and react without significant surface migration. In this instance, the deposition rate is proportional to the arrival angle of the gas molecules. Reactants arriving along the top horizontal surface come from many different angles and ϕ_1, the arrival angle, varies in two dimensions from 0 to 180°, whereas reactants arriving at the top of a vertical wall have an arrival angle ϕ_2 that varies from 0° to 90°. Thus, the film thickness on the top surface is double that on a wall surface. Further down the wall, ϕ_3 is related to the width of the opening, and the film thickness is proportional to

$$\phi_3 \cong \arctan \frac{W}{l}, \tag{19}$$

where l is the distance from the top surface and W is the width of the opening. This type of step coverage is thin along the vertical walls, with a possible crack at the bottom of step caused by self-shadowing.

Silicon dioxide formed by TEOS decomposition at reduced pressure gives a nearly conformal coverage due to rapid surface migration. Similarly, the high-temperature dichlorosilane-nitrous oxide reaction also results in conformal coverage. However, during silane-oxygen deposition, no surface migration takes place and the step coverage is determined by the arrival angle. Most evaporated or sputtered materials have a step coverage similar to that in Fig. 10*b*.

P-Glass Flow

A smooth topography is usually required for the deposited silicon dioxide used as an insulator between metal layers. If the oxide used to cover the lower metal layer is concave, circuit failure may result from an opening that may occur in the upper metal layer during deposition. Because phosphorus-doped silicon dioxide (P-glass) deposited at low temperatures becomes soft and flows upon heating, it provides a smooth surface and is often used to insulate adjacent metal layers. This process is called P-glass flow. In addition, the phosphorus can further getter sodium to prevent its penetration to sensitive gate areas.

Figure 11 shows four cross sections of scanning-electron-microscope photographs of P-glass covering a polysilicon step.[6] All samples are heated in steam at 1100°C for 20 min. Figure 11*a* shows a sample of glass that contains a negligibly small amount of phosphorus and does not flow. Note the concavity of the film and that the corresponding angle θ is about 120°. Figures 11*b*, 11*c*, and 11*d* show samples of P-glass with progressively higher phosphorus contents up to 7.2 wt% (weight percent). In these samples the decreasing step angles of the P-glass layer indicate how flow increases with phosphorus concentration. P-glass flow depends on annealing time, temperature, phosphorus concentration, and the annealing ambient.[6]

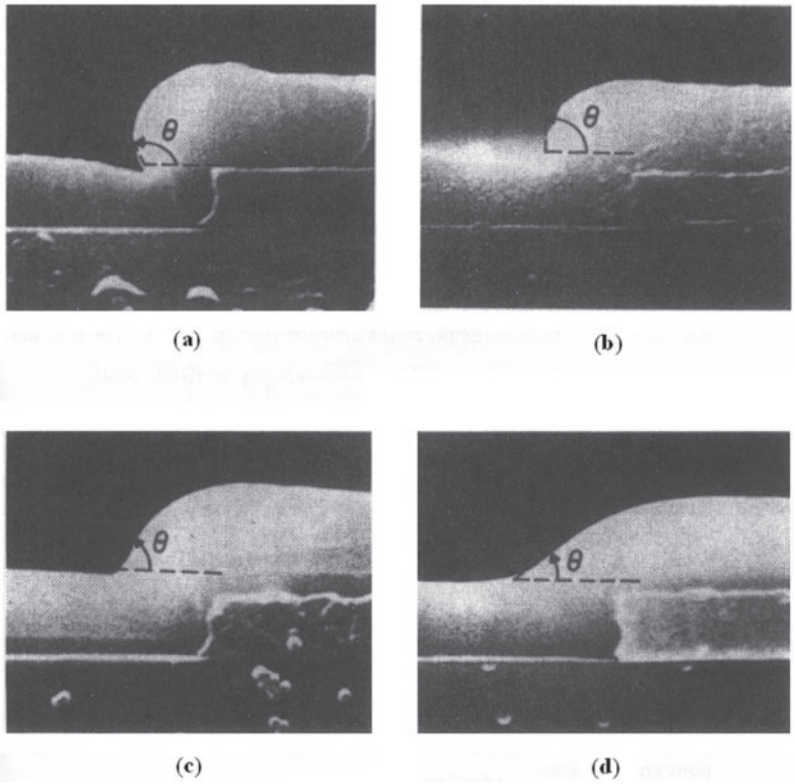


Fig. 11 Scanning-electron-microscope photographs (10,000×) of samples annealed in steam at 1100°C for 20 minutes for the following weight percent phosphorus:6 (*a*) 0 wt%; (*b*) 2.2 wt%; (*c*) 4.6 wt%; and (*d*) 7.2 wt%.

The angle θ as a function of weight percent of phosphorus as shown in Fig. 11 can be approximated by

$$\theta \cong 120^{\circ}\left(\frac{10 \quad wt\%}{10}\right). \tag{20}$$

If we want an angle smaller than 45° we require a phosphorus concentration larger than 6 wt%. However, at concentrations above 8 wt%, the metal film (e.g., aluminum) may be corroded by the acid products formed during the reaction between the phosphorus in the oxide and atmospheric moisture. Therefore, the P-glass flow process uses phosphorus concentrations of 6–8 wt%.

The efficiency of dopant incorporation is controlled by the decomposition mechanism of the dopant sources. In the thermal process, temperature is the dominant factor. In the plasma-enhanced process, the temperature dependence is much less, and plasma power is much more critical.

12.2.3 Silicon Nitride

It is difficult to grow silicon nitride by thermal nitridation (e.g., with ammonia, NH_3) because of its low growth rate and high growth temperature. However, silicon nitride films can be deposited by an intermediate-temperature (750°C) LPCVD process or a low-temperature (300°C) plasma-assisted CVD process.[7,8] The LPCVD films are of stoichiometric composition (Si_3N_4) with high density (2.9–3.1 g/cm^3). These films can be used to passivate devices because they serve as good barriers to the diffusion of water and sodium. The films also can be used as masks for the selective oxidation of silicon because silicon nitride oxidizes very slowly and prevents the underlying silicon from oxidizing. The films deposited by plasma-assisted CVD are not stoichiometric and have a lower density (2.4–2.8 g/cm^3). Because of the low

deposition temperature, silicon nitride films can be deposited over fabricated devices and serve as their final passivation. The plasma-deposited nitride provides excellent scratch protection, serves as a moisture barrier, and prevents sodium diffusion.

In the LPCVD process, dichlorosilane and ammonia react at reduced pressure to deposit silicon nitride at temperatures between 700° and 800°C. The reaction is

$$3SiCl_2H_2 + 4NH_3 \xrightarrow{\sim 750^0\,C} Si_3N_4 + 6HCl + 6H_2\,. \tag{21}$$

Good film uniformity and high wafer throughput (the number of wafers processed per hour) are advantages of the reduced-pressure process. As in oxide deposition, silicon nitride deposition is controlled by temperature, pressure, and reactant concentration. The activation energy for deposition is about 1.8 eV. The deposition rate increases with increasing total pressure or dichlorosilane partial pressure and decreases with increasing ammonia-to-dichlorosilane ratio.

Silicon nitride deposited by LPCVD is an amorphous dielectric containing up to 8 atomic percent (at%) hydrogen. The etch rate in buffered HF is less than 1 nm/min. The film has a very high tensile stress of approximately 10^{10} dynes/cm^2, which is nearly 10 times that of TEOS-deposited SiO_2. Films thicker than 200 nm may crack because of the very high stress. The resistivity of silicon nitride at room temperature is about 10^{16} Ω-cm. Its dielectric constant is 7 and its dielectric strength is 10^7 V/cm.

In the plasma-assisted CVD process, silicon nitride is formed either by reacting silane and ammonia in an argon plasma or by reacting silane in a nitrogen discharge. The plasma dissociates the precursors and creates high-energy forms of the reactant species that accelerate the reaction rate at a much lower temperature. Ions and electrons are charged species associated with plasma. The reactions are as follows:

$$SiH_4 + NH_3 \xrightarrow{300^0\,C} SiNH + 3H_2, \tag{22a}$$

$$2SiH_4 + N_2 \xrightarrow{300^0\,C} 2SiNH + 3H_2. \tag{22b}$$

The products depend strongly on deposition conditions. The radial-flow pallel-plate reactor (Fig. 9*b*) is used to deposit the films. The deposition rate generally increases with increasing temperature, power input, and reactant gas pressure.

Large concentrations of hydrogen are contained in plasma-deposited films. The plasma nitride (also referred to as SiN) used in semiconductor processing generally contains 20-25 at% hydrogen. Films with low tensile stress (~2 × 10^9 dynes/cm^2) can be prepared by plasma deposition. Film resistivities range from 10^5 to 10^{21} Ω-cm, depending on silicon-to-nitrogen ratio, whereas dielectric strengths are between 1 × 10^6 and 6 × 10^6 V/cm. For passivation, the films must be a moisture and sodium diffusion barrier with good step coverage and no pinholes. Silicon nitride is an ideal material for a passivation layer, but high-temperature thermally deposited nitride exceeds the temperature for Al metallization and the hydrogen content in lower temperature PECVD nitride can cause a degradation in hot carrier lifetime.

12.2.4 Low-Dielectric-Constant Materials

As devices continue to scale down to the deep submicron region, they require multilevel interconnection architecture to minimize the time delay due to parasitic resistance *R* and capacitance *C*. The gain in device speed at the gate level will be offset by the propagation delay at the metal interconnects because of the increased *RC* time constant, as shown in Fig. 12. For example, in devices with gate length of 250 nm or less, up to 50% of the time delay is due to the *RC* delay of long interconnections.[9] Therefore, the device interconnection network becomes a limiting factor in determining chip performance by affecting device speed, cross talk, and power consumption of ULSI circuits.

To reduce the *RC* time constant of ULSI circuits, interconnection materials with low resistivity and interlayer films with low capacitance are required. For low-capacitance ($C = \varepsilon_i A/d$, where ε_i is the dielectric permittivity, A the area, and d the thickness of the dielectric film), it is not easy to lower the parasitic capacitance by increasing the thickness d of the interlayer dielectric (which makes gap filling more difficult), or decreasing wiring height and area A (which results in the increase of interconnect resistance). Therefore, materials with a low dielectric constant (low k) are required. The ε_i is equal to the product of k and ε_o, where k and ε_o are the dielectric constant and the vacuum permittivity, respectively.

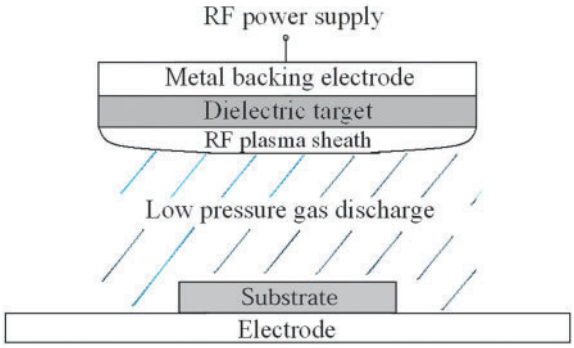


Fig. 21 **Schematic diagram of RF sputtering.**

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:

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

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

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

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:

$$6TiCl_4 + 8NH_3 \rightarrow 6\ TiN + 24HCl + N_2, \quad (27)$$

$$2TiCl_4 + N_2 + 4H_2 \rightarrow 2TiN + 8HCl, \quad (28)$$

$$2TiCl_4 + 2NH_3 + H_2 \rightarrow 2\ TiN + 8HCl. \quad (29)$$

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

- 14.1 BASIC DIFFUSION PROCESS
- 14.2 EXTRINSIC DIFFUSION
- 14.3 DIFFUSION-RELATED PROCESSES
- 14.4 RANGE OF IMPLANTED IONS
- 14.5 IMPLANT DAMAGE AND ANNEALING
- 14.6 IMPLANTATION-RELATED PROCESSES
- SUMMARY

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.

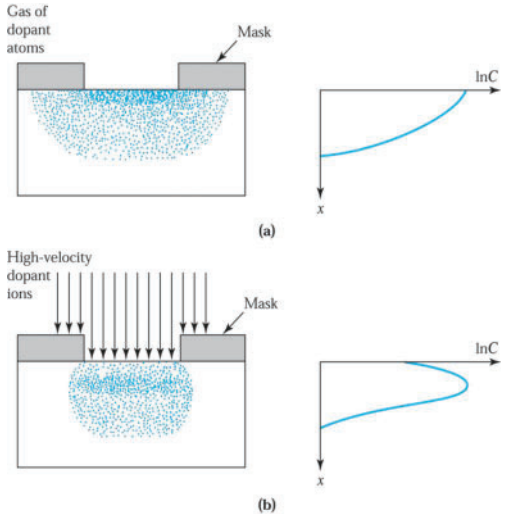


Fig. 1 **Comparison of (*a*) diffusion and (*b*) ion-implantation techniques for selective introduction of dopants into the semiconductor substrate.**

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.