

电子科学与技术专业英语

微电子技术分册

总主编 张爱红

主编 陈伟平 田丽

哈尔滨工业大学出版社

专业英语阅读丛书

● 电子科学与技术专业英语

全套 48.00 元

光电子技术分册

微电子技术分册

光电信息技术分册

对外贸易英文函电

12.00 元

国际商贸英语

15.00 元

力学专业英语

12.00 元

信息功能材料专业英语

13.80 元

土木工程专业英语 (第二版)

12.80 元

机电工程专业英语 (第三次修订)

12.00 元

机械工程专业英语 (第三版)

13.80 元

热能与动力工程专业英语

14.00 元

电子信息与通信专业英语

12.00 元

环境科学与工程专业英语

12.00 元

材料科学与工程专业英语 (修订版)

13.80 元

计算机专业英语 (修订版)

11.00 元

建筑工程专业英语

11.50 元

电气自动化专业英语

13.80 元

外贸英语阅读

8.50 元

汽车英语阅读

17.50 元

会计英语阅读

12.80 元

经济专业英语阅读

14.80 元

自动控制专业英语阅读

13.80 元

应用化学专业英语

17.00 元

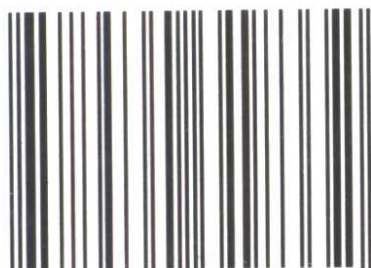
管理类专业英语阅读

估 12.80 元

责任编辑 张晓京

封面设计 卞秉利

ISBN 7-5603-1833-9



9 787560 318332 >

ISBN 7-5603-1833-9 / TN·65

全套(三册)定价 48.00 元

电子科学与技术专业英语

微电子技术分册

总主编 张爱红

主 编 陈伟平 田 丽

哈尔滨工业大学出版社

哈尔滨

内 容 简 介

本书由半导体物理、半导体器件、半导体与集成电路工艺、集成电路设计、微电子机械系统和科技文献范例六部分组成。全书涵盖了微电子技术领域的基本内容,同时又介绍了该领域的一些较新的发展。

本书可作为电子科学与技术专业微电子技术方向的三、四年级本科生的专业英语教材,也可供相关专业的本科生和从事该专业领域的工程技术人员使用。

图书在版编目(CIP)数据

电子科学与技术专业英语.微电子技术分册/张爱红主编.—哈尔滨:哈尔滨工业大学出版社,2003.6

ISBN 7-5603-1833-9

I.电… II.张… III.①电子技术-英语-高等学校-教材
②微电子技术-英语-高等学校-教材 IV.H31

中国版本图书馆 CIP 数据核字(2003)第 037940 号

出版发行 哈尔滨工业大学出版社
社 址 哈尔滨市南岗区教化街 21 号 邮编 150006
传 真 0451-6414749
印 刷 地矿部黑龙江测绘印制中心印刷厂
开 本 850×1168 1/32 印张 12.75 字数 326 千字
版 次 2003 年 6 月第 1 版 2003 年 6 月第 1 次印刷
书 号 ISBN 7-5603-1833-9/TN·65
印 数 1~3 000
定 价 全套(三册)定价 48.00 元 本册定价 16.00 元

前 言

21 世纪是国际化的高科技时代,国际化的信息时代,作为国际间交流的重要载体——英语,其作用显得更为重要。专业英语是大学外语教学的重要组成部分,其目的是强化巩固基础英语并进行实践应用,从而掌握科技英语技能,使学生能够熟练阅读国外相关文献,掌握国内外本专业发展前沿的最新动态,并具有一定的科技写作能力。为此我们编写这套专业英语教材,以满足高等院校电子科学与技术 and 电子信息科学与技术专业及相关专业的本科生专业英语教学的需要以及从事上述专业的工程技术人员学习英语的要求。全套教材分三册,包括光电子技术分册、微电子技术分册和光电信息技术分册。

本书是微电子技术分册,由陈伟平副教授担任主编,参编人员有田丽、王东红、叶水驰、王进祥、来逢昌。刘晓为教授担任主审。

本分册内容分为五部分:第一部分为半导体物理的基本理论。主要内容包括:半导体、晶格、能带、载流子等基本概念;硅及砷化镓物理学特性;P-N 结及其特性。第二部分为半导体器件。主要内容包括:双极型器件;MOS 器件;微波与光电子器件。第三部分为半导体工艺。主要内容包括晶体生长、薄膜生长、光刻、腐蚀、扩散、注入等基本半导体工艺;双极型集成电路工艺;MOS 集成电路工艺。第四部分为集成电路设计。主要内容包括:设计方法;综合与仿真。第五部分为 MEMS。主要内容包括:MEMS 概念与特点;MEMS 工艺、器件与材料等。第六部分为科技文献范例。主要介绍微电子技术的现状与未来发展。

书中内容选材精良,但考虑到篇幅的限制及上下文的连贯性,对所选内容的图表及文字进行了部分删减和改动。为了有利于教

学和阅读理解,帮助学生更好地理解原文,每章后面给出了主要专业词汇及难句注释,并在每部分后面列出了有关参考文献。

本书是根据编者在专业英语方面的长期教学经验和体会编写而成的,内容较多,教师可以根据实际教学时数和需要选取合适的章节作为课堂教学内容,其他部分可作为学生课后阅读材料。对于其他专业的读者,也可通过阅读本书对微电子学原理及其应用有一定的了解和认识,既学习了英语,又开阔了视野。

本书在编写过程中得到电子科学与技术系的许多老师和学生的大力支持,在此一并表示诚挚的感谢。由于编者水平有限,书中难免还存在一些缺点和疏漏,殷切希望广大读者批评指正。

编 者

2003.5

CONTENTS

Chapter 1 Semiconductors Physics	1
1.1 Energy Bands and Carrier Concentration	1
1.2 Carrier Transport Phenomena	17
1.3 PN Junction	29
1.4 Summary	48
参考文献	72
Chapter 2 Semiconductor Device	73
2.1 Bipolar Junction Transistor	73
2.2 The MOSFET	98
2.3 Microwave and Photonic Devices	122
2.4 Summary	129
参考文献	156
Chapter 3 Processing Technology	157
3.1 Crystal Growth and Epitaxy	157
3.2 Crystal Growth from the Melt	158
3.3 Vapor-Phase Epitaxy	162
3.4 Oxidation and Film Deposition	165
3.5 Diffusion and Ion Implantation	171
3.6 Lithographies	178
3.7 Wet Chemical Etching	181
3.8 Dry Etching	185
3.9 Integrated Devices	188
3.10 Fundamental Limits of Integrated Devices	215
参考文献	223

Chapter 4 Integrated Circuits	224
4.1 Introduction	224
4.2 Design Analysis and Simulation	225
4.3 Verification	248
4.4 Summary	270
参考文献	272
Chapter 5 Microelectromechanical Systems (MEMS)	273
5.1 Introduction	273
5.2 MEMS	276
5.3 Mechanical Characteristics of Silicon	282
5.4 Microfabrications for MEMS	289
5.5 Microsensing for MEMS	306
5.6 Electromechanical Actuation	311
5.7 Materials for MEMS	312
参考文献	317
Chapter 6 Examples of Scientific and Technological Papers	
.....	318
6.1 The Challenges for Physical Limitations in Si Microelectronics	
.....	318
6.2 Microelectronics and Photonics – the Future	329
6.3 Accelerated Verification of Digital Devices Using VHDL	
.....	368
6.4 What Is Nanotechnology	390
参考文献	401

Chapter 1

Semiconductors Physics

1.1 Energy Bands and Carrier Concentration

1.1.1 Semiconductor Materials

Solid-state materials can be grouped into three classes—insulators, semiconductors, and conductors. Figure 1 – 1 shows the electrical conductivities σ (and the corresponding resistivities $\rho \equiv 1/\sigma$) associated with some important materials in each of the three classes. Insulators such as fused quartz and glass have very low conductivities, in the order of 1E-18 to 1E-8 S/cm; and conductors such as aluminum and silver have high conductivities, typically from 10^4 to 10^6 S/cm. Semiconductors have conductivities between those of insulators and those of conductors. The conductivity of a semiconductor is generally sensitive to temperature, illumination, magnetic field, and minute amount of impurity atoms. This sensitivity in conductivity makes the semiconductor one of the most important materials for electronic applications.

The study of semiconductor materials began in the early nineteenth century. Over the years many semiconductors have been investigated. Table 1 shows a portion of the periodic table

related to semiconductors. The element semiconductors, those composed of single species of atoms, such as silicon (Si) and germanium (Ge), can be found in Column IV. However, numerous compound semiconductors are composed of two or more elements. For example, gallium arsenide (GaAs) is a III-V compound that is a combination of gallium (Ga) from Column III and arsenic (As) from Column V.

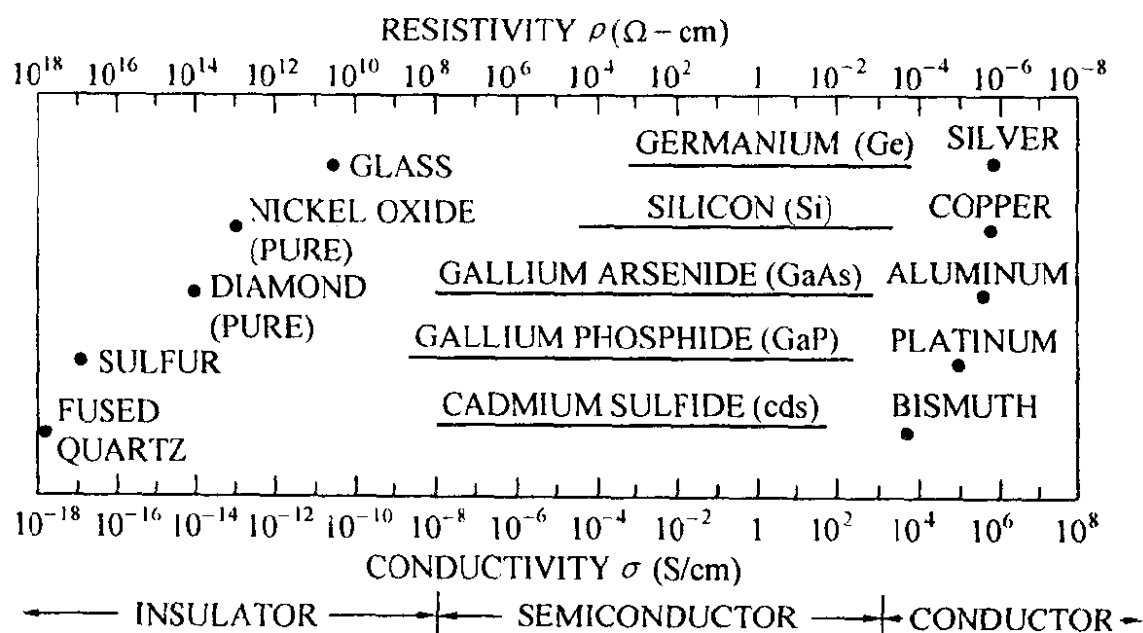


Figure 1 - 1 Typical range of conductivities for insulators, semiconductors, and conductors

Table 1 - 1 Portion of the Periodic Table Related to Semiconductors

Period	Column I	III	IV	V	VI
2		B Boron	C Carbon	N Nitrogen	
3	Mg Magnesium	Al Aluminum	Si Silicon	P Phosphorus	S Sulfur
4	Zn Zinc	Ga Gallium	Ge Germanium	As Arsenic	Se Selenium
5	Cd Cadmium	In Indium	Sn Tin	Sb Antimony	Te Tellurium
6	Hg Mercury		Pb Lead		

Prior to the invention of the bipolar transistor in 1947, semiconductors were used only as two-terminal devices, such as rectifiers and photodiodes. In the early 1950s, germanium was the major semiconductor material. However, germanium proved unsuitable in many applications because germanium devices exhibited high leakage currents at only moderately elevated temperatures. In addition, germanium oxide is water soluble and unsuited for device fabrication. Since the early 1960s silicon has become a practical substitute and has now virtually supplanted germanium as a material for semiconductor fabrication. The main reasons we now use silicon are that silicon devices exhibit much lower leakage currents, and high-quality silicon dioxide can be grown thermally. There is also an economic consideration. Device grade silicon costs much less than any other semiconductor material. Silicon in the form of silica and silicates comprises 25 % of the Earth's crust, and silicon is second only to oxygen in abundance. At present, silicon is one of the most studied elements in the periodic table; and silicon technology is by far the most advanced among all semiconductor technologies.

Many of the compound semiconductors have electrical and optical properties that are absent in silicon. These semiconductors, especially gallium arsenide (GaAs), are used mainly for microwave and photonic applications. Although we do not know as much about the technology of compound semiconductors as we do about that of silicon, compound semiconductor technology has advanced partly because of the advances in silicon technology. In this book we are concerned mainly with device physics and processing technology of silicon and gallium arsenide.

1.1.2 Crystal Structure

The semiconductor materials we will study are single crystals, that is, the atoms are arranged in a three-dimensional periodic fashion. The periodic arrangement of atoms in a crystal is called a lattice. In a crystal, an atom never strays far from a single, fixed position. The thermal vibrations associated with the atom are centered about this position. For a given semiconductor, there is a unit cell that is representative of the entire lattice; by repeating the unit cell throughout the crystal, one can generate the entire lattice.

Figure 1 – 2 shows some basic cubic-crystal unit cells. Figure 1 – 2(a) shows a simple cubic crystal; each corner of the cubic lattice is occupied by an atom that has six equidistant nearest neighboring atoms. The dimension a is called the lattice constant. Only polonium is crystallized in the simple cubic lattice. Figure 1 – 2(b) is a body-centered cubic(bcc) crystal, where in addition to the eight corner atoms, an atom is located at the center of the cube. In a bcc lattice, each atom has eight nearest-neighboring atoms. Crystals exhibiting bcc lattices include those of sodium and tungsten. Figure 1 – 2(c) shows a face-centered cubic (fcc) crystal that has one atom at each of the six cubic faces in addition to the eight corner atoms. In an fcc lattice, each atom has 12 nearest neighboring atoms. A large number of elements exhibit the fcc lattice form, including aluminum, copper, gold, and platinum.

The element semiconductors, silicon and germanium, have a diamond lattice structure. This structure also belongs to the cubic-crystal family and can be seen as two interpenetrating fcc

sublattices with one sublattice displaced from the other by one quarter of the distance along a diagonal of the cube (i. e., a displacement of a $\sqrt{3}/4$). All atoms are identical in a diamond lattice, and each atom in the diamond lattice is surrounded by four equidistant nearest neighbors that lie at the corners of a tetrahedron. Most of the III-V compound semiconductors (e. g., GaAs) have a zincblende lattice, which is identical to a diamond lattice except that one fcc sublattice has Column III atoms (Ga) and the other has Column V atoms (As).

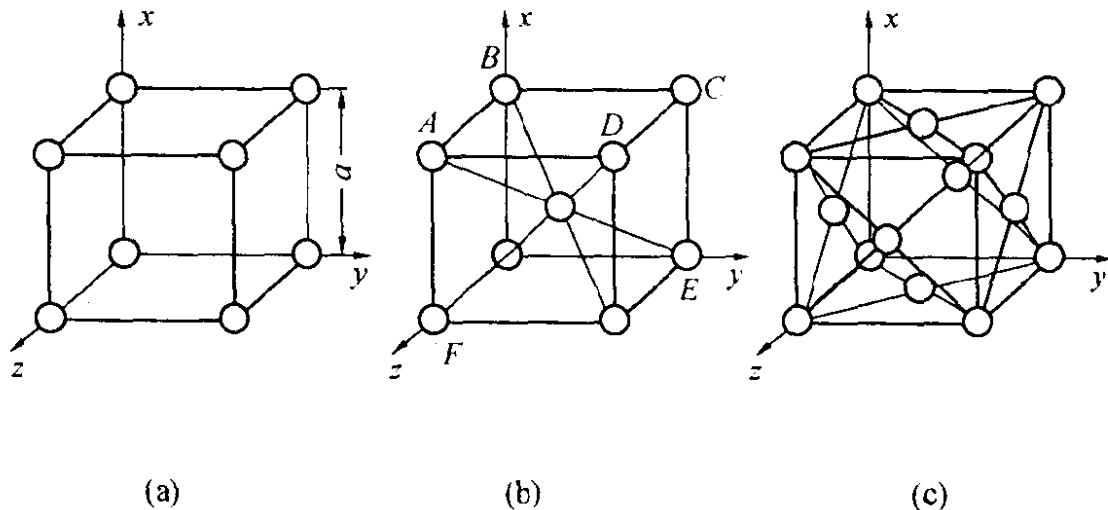


Figure 1 - 2 Three cubic-crystal unit cells

Therefore, the crystal properties along different planes are different, and the electrical and other device characteristics are dependent on the crystal orientation. A convenient method of defining the various planes in a crystal is to use Miller indices. These indices are obtained using the following steps:

- (1) Find the intercepts of the plane on the three Cartesian coordinates in terms of the lattice constant.
- (2) Take the reciprocals of these numbers and reduce them to the smallest three integers having the same ratio.
- (3) Enclose the result in parentheses (*hkl*) as the Miller

indices for a single plane.

Some conventions are given as follows:

(hkl) : For a plane that intercepts the x -axis on the negative side of the origin, such as $(\bar{1}00)$.

$\{hkl\}$: For planes of equivalent symmetry-such as $\{100\}$ for (100) , (010) , (001) , $(\bar{1}00)$, $(0\bar{1}0)$, and $(00\bar{1})$ in cubic symmetry.

$[hkl]$: For a crystal direction, such as $[100]$ for the x -axis. Thus, the $[100]$ -direction is perpendicular to (100) -plane, and the $[111]$ -direction is perpendicular to the (111) -plane.

$\langle hkl \rangle$: For a full set of equivalent directions-such as $\langle 100 \rangle$ for $[100]$, $[010]$, $[001]$, $[\bar{1}00]$, $[0\bar{1}0]$, and $[00\bar{1}]$.

1.1.3 Valence Bonds

As discussed in Section 1.1.2, each atom in a diamond lattice is surrounded by four nearest neighbors. Each atom has four electrons in the outer orbit, and each atom shares these valence electrons with its four neighbors. This sharing of electrons is known as covalent bonding; each electron pair constitutes a covalent bond. Covalent bonding occurs between atoms of the same element, or between atoms of different elements that have similar outer-shell electron configurations. Each electron spends an equal amount of time with each nucleus. However, both electrons spend most of their time between the two nuclei. The force of attraction for the electrons by both nuclei holds the two atoms together. For a zincblende lattice such as gallium arsenide, the major bonding force is from the covalent bonds. However, gallium arsenide has a slight ionic bonding force that is an

electrostatic attractive force between each Ga^- ion and its four neighboring As^+ ions, or between each As^+ ion and its four neighboring Ga^- ions.

At low temperatures, the electrons are bound in their respective tetrahedron lattice; consequently, they are not available for conduction. At higher temperatures, thermal vibrations may break the covalent bonds. When a bond is broken, a free electron results that can participate in current conduction. An electron deficiency is left in the covalent bond. This deficiency may be filled by one of the neighboring electrons, which results in a shift of the deficiency location, as from location A to location B. We may therefore consider this deficiency as a particle similar to an electron. This fictitious particle is called a hole. It carries a positive charge and moves, under the influence of an applied electric field, in the direction opposite to that of an electron. The concept of a hole is analogous to that of a bubble in a liquid. Although it is actually the liquid that moves, it is much easier to talk about the motion of the bubble in the opposite direction.

1.1.4 Energy Bands

For an isolated atom, the electrons of the atom can have only discrete energy levels. For example, the energy levels for an isolated hydrogen atom are given by the Bohr model

$$E_h = - m_0 q^4 / (8 \epsilon_0^2 h^2 n^2) = - 13.6 / n^2 \text{ eV}$$

where m_0 is the free-electron mass, q is the electronic charge, ϵ_0 is the free- space permittivity, h is the Planck constant, and n is a positive integer called the principal quantum number. The discrete energies are $- 13.6 \text{ eV}$ for the ground level ($n = 1$), $- 3.4 \text{ eV}$ for the first excited level ($n = 2$), etc

We now consider two identical atoms. When they are far apart, the allowed energy levels for a given principal quantum number (e.g., $n = 1$) consist of one doubly degenerate level, that is, each atom has exactly the same energy (e.g., -13.6 eV for $n = 1$). As the two atoms approach one another, the doubly degenerate energy level will split into two levels by the interaction between the atoms. When we bring N atoms together to form a crystal, the N -fold degenerate energy level will split into N separate but closely spaced levels due to atomic interaction. This results in an essentially continuous band of energy.

The detailed energy band structures of crystalline solids have been calculated using quantum mechanics. Figure 1 – 3 is a schematic diagram of the formation of a diamond lattice crystal from isolated silicon atoms. Each isolated atom has its discrete energy levels (two levels are shown on the far right of the diagram). As the interatomic spacing decreases, each degenerate energy level splits to form a band. Further decrease in spacing causes the bands originating from different discrete levels to lose their identities and merge together, forming a single band. When the distance between atoms approaches the equilibrium interatomic spacing of the diamond lattice (with a lattice constant of 0.357 nm for silicon), this band splits again into two bands. These bands are separated by a region which designates energies that the electrons in the solid cannot possess. This region is called the forbidden gap, or bandgap E_g . The upper band is called the conduction band, while the lower band is called the valence band, as shown on the far left of Figure 1 – 3.

Figure 1 – 4 shows the energy band diagrams of three classes of solids—insulators, semiconductors, and conductors. In an

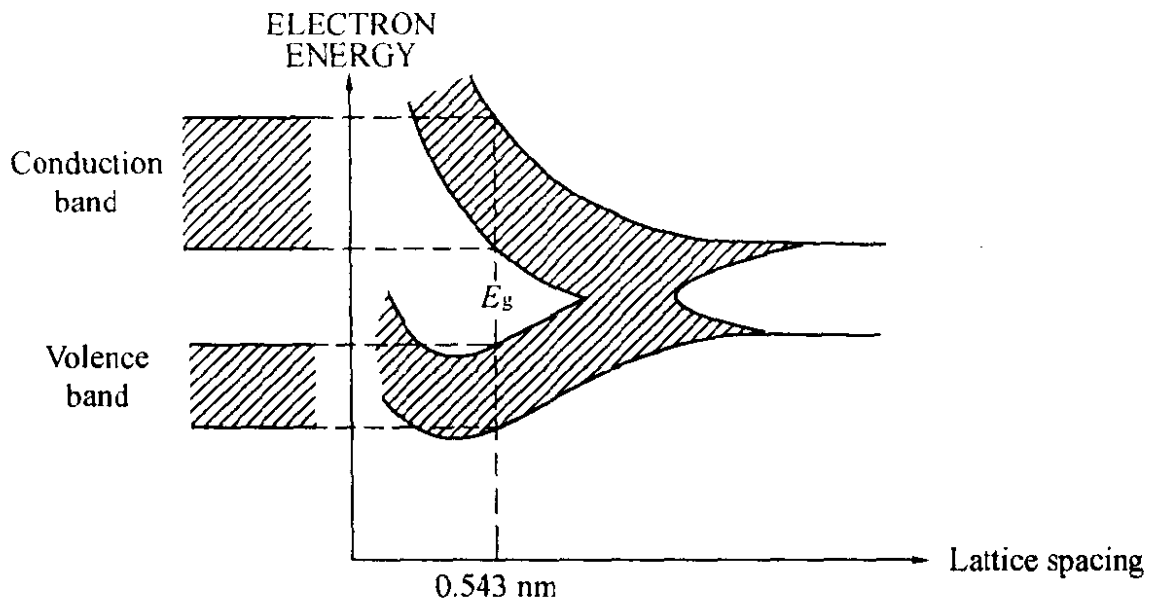


Figure 1 – 3 Formation of energy bands

insulator such as silicon dioxide (SiO_2), the valence electrons form strong bonds between neighboring atoms. These bonds are difficult to break, and consequently there are no free electrons to participate in current conduction. As shown in the energy band diagram Figure 1 – 4(a), there is a large bandgap. Note that all energy levels in the valence band are occupied by electrons and all energy levels in the conduction band are empty. Thermal energy or an applied electric field cannot raise the uppermost electron in the valence band to the conduction band. Therefore, silicon dioxide is an insulator, which cannot conduct current.

As we discussed in Section 1.1.3, bonds between neighboring atoms in a semiconductor are only moderately strong. Therefore, thermal vibrations will break some bonds. When a bond is broken, a free electron along with a free hole result. Figure 1 – 4(b) shows that the bandgap of a semiconductor is not as large as that of an insulator (e.g., Si with a bandgap of 1.12 eV). Because of this, some electrons will be able to move from the valence band to the conduction band, leaving holes in the valence band. When an

electric field is applied, both the electrons in the conduction band and the holes in the valence band will gain kinetic energy and conduct electricity.

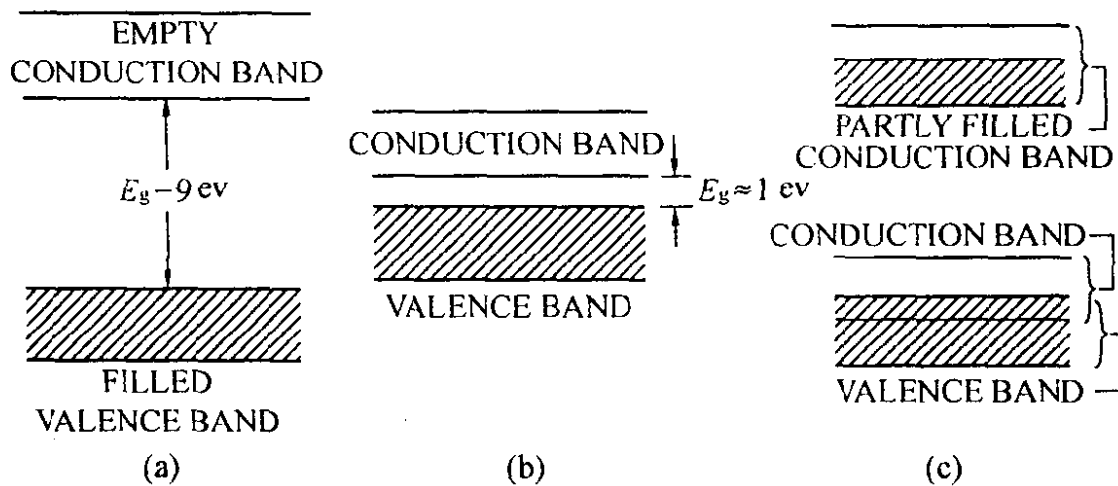


Figure 1 - 4 Schematic energy band representations

In conductors such as metals, Figure 1 - 4 (c), the conduction band either is partially filled or overlaps the valence band so that there is no bandgap. As a consequence, the uppermost electrons in the partially filled band or electrons at the top of the valence band can move to the next-higher available energy level when they gain kinetic energy (e.g., from an applied electric field). Therefore, current conduction can readily occur in conductors.

The energy band diagrams shown in Figure 1 - 4 indicate electron energies. When the energy of an electron is increased, the electron moves to a higher position in the band diagram. On the other hand, when the energy of a hole is increased, the hole moves downward in the valence band. (This is because a hole has a charge opposite that of an electron.)

As we discussed before, the separation between the energy of the lowest conduction band and that of the highest valence band is

called the bandgap E_g , which is the most important parameter in semiconductor physics. We designate E_c as the bottom of the conduction band; E_c corresponds to the potential energy of an electron, that is, the energy of a conduction electron that is at rest. The kinetic energy of an electron is measured upward from E_c . Similarly, we designate E_v as the top of the valence band; E_v corresponds to the potential energy of a hole. The kinetic energy of a hole is measured downward from E_v .

At room temperature and under normal atmosphere, the values of the bandgap are 1.12 eV for silicon and 1.42 eV for gallium arsenide. The bandgap approaches 1.17 eV for silicon and 1.52 eV for gallium arsenide at 0 K.

1.1.5 Density of States

When electrons move back and forth along the x -direction in a semiconductor material, the movements can be described by standing-wave oscillations. The wavelength λ of a standing wave is related to the length of the semiconductor L by

$$\frac{L}{\lambda} = n_x \quad (1-1)$$

where n_x , is an integer. The wavelength can be expressed as

$$\lambda = \frac{h}{p_x} \quad (1-2)$$

where h is the Planck constant and $\bar{p}_x$, is the crystal momentum (动量) in the x -direction. Substituting Eq. 1 - 2 into Eq. 1 - 1 gives

$$L\bar{p}_x = hn_x \quad (1-3)$$

The incremental momentum $d\bar{p}$ required for a unity increase in n_x , is

$$L d\bar{p}_x = h \quad (1-4)$$

For a three-dimensional cube of side L , we have

$$L^3 d\bar{p}_x d\bar{p}_y d\bar{p}_z = h^3 \quad (1-5)$$

The volume $d\bar{p}_x d\bar{p}_y d\bar{p}_z$ in the momentum space for a unit cube ($L = 1$) is thus equal to h^3 . Each incremental change in n corresponds to a unique set of *integers* (n_x, n_y, n_z), which in turn corresponds to an allowed energy state. Thus, the volume in momentum space for an energy state is h^3 . The volume between two concentric spheres (from $\bar{p}$ to $\bar{p} + d\bar{p}$) is $4\pi\bar{p}^2 d\bar{p}$. The number of energy states contained in this volume is then $2(4\pi\bar{p}^2 d\bar{p})/h^3$, where the factor 2 accounts for the electron spins. We can substitute E for $\bar{p}$ and obtain

$$N(E) dE = \frac{8\pi\bar{p}^2 d\bar{p}}{h^3} = 4\pi \left[\frac{2m_n}{h^2} \right]^{3/2} E^{1/2} dE \quad (1-6)$$

and

$$N(E) = 4\pi \left[\frac{2m_n}{h^2} \right]^{3/2} E^{1/2} \quad (1-7)$$

where $N(E)$ is called the density of states, that is, the density of allowed energy states per unit volume.

1.1.6 Intrinsic Carrier Concentration

At finite temperatures, continuous thermal agitation results in the excitation of electrons from the valence band to the conduction band and leaves an equal number of holes in the valence band. An intrinsic (本征) semiconductor is one that contains relatively small amounts of impurities (杂质) compared to the thermally generated electrons and holes. To obtain the electron density (i. e., the number of electrons per unit volume) in an intrinsic semiconductor, we first evaluate the electron density in an

incremental energy range dE . This density $n(E)$ is given by the product of the density of allowed energy states per unit volume $N(E)$ and by the probability of occupying that energy range $F(E)$. Thus, the electron density in the conduction band is given by integrating $N(E) F(E) dE$ from the bottom of the conduction band (E_c initially taken to be $E = 0$ for simplicity) to the top of the conduction band E_{top}

$$n = \int_0^{E_{\text{top}}} n(E) dE = \int_0^{E_{\text{top}}} N(E) F(E) dE \quad (1-8)$$

The probability that an electronic state with energy E is occupied by an electron is given by the *Fermi-Dirac distribution function* (which is also referred to as the Fermi distribution function)

$$F(E) = \frac{1}{1 + \exp(E - E_F)/kT} \quad (1-9)$$

where k is the Boltzmann constant, T is the absolute temperature in degrees Kelvin, and E_F is the Fermi level. The Fermi level is the energy at which the probability of occupation by an electron is exactly one half. The Fermi distribution function can thus be approximated by simpler expressions

$$\begin{aligned} F(E) &\approx \exp(E - E_F)/kT \\ \text{for } (E - E_F) &> 3kT \end{aligned} \quad (1-10a)$$

and

$$\begin{aligned} F(E) &\approx 1 - \exp[-(E_F - E)/kT] \\ \text{for } (E - E_F) &< -3kT \end{aligned} \quad (1-10b)$$

Equation 1-10b can be regarded as the probability that a hole occupied a state located at energy E .

Figure 1-5 shows schematically from left to right the band diagram, the density of states (which varies as $\sqrt{E}$), the Fermi distribution function, and the carrier concentrations for an intrinsic

semiconductor. The carrier concentrations can be obtained graphically from Figure 1 – 5 using Eq. 13; that is, the product of $N(E)$ in Figure 1 – 5(b) and $F(E)$ in Figure 1 – 5(c) gives the $n(E)$ -versus- E curve(upper curve) in Figure 1 – 5(d). The upper shaded area in Figure 1 – 5d corresponds to the electron density.

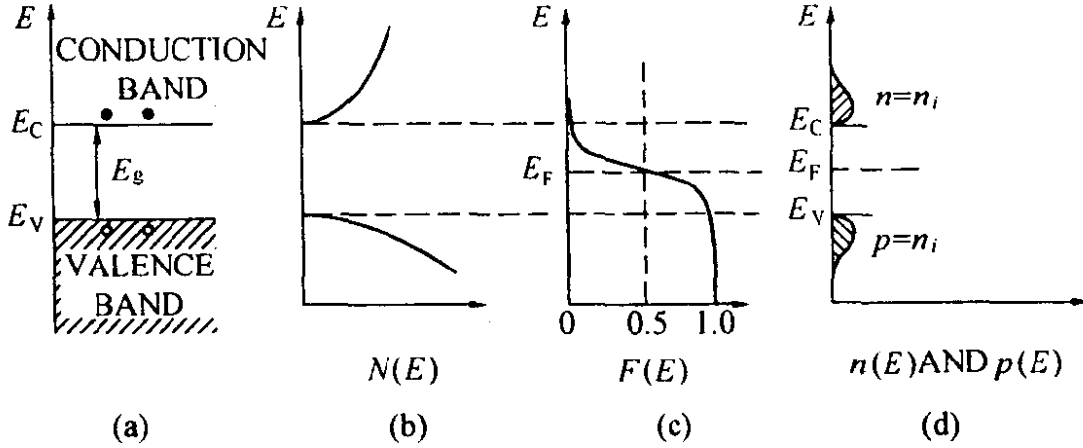


Figure 1 – 5 Intrinsic semiconductor

In the conduction band there are a large number of allowed states. However, for intrinsic semiconductors there will be only a few electrons in the conduction band, hence the probability of an electron occupying one of these states is small. There also are a large number of allowed states in the valence band. By contrast most of those are occupied by electrons. Thus, the probability of an electron occupying one of these states in the valence band is nearly unity. There will be only a few unoccupied electron states, that is, holes, in the valence band. As can be seen, the Fermi level is located near the middle of the bandgap (i.e., E_F is many kT below E_C). Because $F(E)$ is an exponentially decreasing function of E (Eq. 1 – 10a), the value of E_{top} can be replaced by infinity. Substituting Eq. 1 – 10(a) into Eq. 1 – 8 yields

$$n = 4\pi \left[\frac{2m_n}{h^2} \right]^{3/2} \int_0^\infty E^{1/2} \exp \left[- \frac{E - E_F}{kT} \right] dE \quad (1 - 11)$$

and
$$N_c \equiv 2 \left[\frac{2\pi m_n kT}{h^2} \right]^{3/2}$$

we obtain for the electron density in the conduction band where N_c is the effective density of states in the conduction band; At room temperature (300 K), N_c is $2.8 \text{ E}19 \text{ cm}^{-3}$ for silicon and $4.7 \text{ E}17 \text{ cm}^{-3}$ for gallium arsenide.

Similarly, we can obtain the hole density p in the valence band.

$$N_v = 2 \left[\frac{2\pi m_p kT}{h^2} \right]^{3/2}$$

where N_v is the effective density of states in the valence band. At room temperature, N_v is $1.04 \text{ E}19 \text{ cm}^{-3}$ for silicon and $7.0 \text{ E}18 \text{ cm}^{-3}$ for gallium arsenide.

For an intrinsic semiconductor as defined previously, the number of electrons per unit volume in the conduction band is equal to the number of holes per unit volume in the valence band, that is, $n = p = n_i$, where n_i is the intrinsic carrier density. This relation of electrons and holes is depicted in Figure 1 – 5(d). Note that the shaded area in the conduction band is the same as that in the valence band.

At room temperature, the second term is much smaller than the bandgap. Hence, the intrinsic Fermi level E_i , that is, the Fermi level of an intrinsic semiconductor, generally lies very close to the middle of the bandgap.

The intrinsic carrier density is obtained

$$n_i = \sqrt{N_c N_v} \exp \left[- \frac{E_g}{kT} \right] \quad (1 - 12)$$

$$np = n_i^2 \quad (1 - 13)$$

where $E_g = (E_c - E_v)$. Equation 1 – 13 is called the mass action

law, which is valid for both intrinsic and extrinsic (i.e., doped with impurities) semiconductors under a thermal equilibrium condition. In an extrinsic semiconductor, the increase of one type of carrier tends to reduce the number of the other type through recombination; thus, the product of the two types of carriers will remain constant at a given temperature.

1.1.7 Donors and Acceptors

When a semiconductor is doped with impurities, the semiconductor becomes *extrinsic* and impurity energy levels are introduced. A silicon atom is replaced (or substituted) by an arsenic atom with five valence electrons. The arsenic atom forms covalent bonds with its four neighboring silicon atoms. The fifth electron becomes a conduction electron that is “donated” to the conduction band. The silicon becomes n-type because of the addition of the negative charge carrier, and the arsenic atom is called a donor. Similar, when a boron atom with three valence electrons substitute for a silicon atom, additional electron is accepted to form four covalent bonds around the boron, and a positively charged “hole” is created in the valence band. This is a p-type semiconductor, and the boron is an acceptor.

For shallow donors in silicon and gallium arsenide, there usually is enough thermal energy to supply the energy E_D to ionize all donor impurities at room temperature and thus provide an equal number of electrons in the conduction band. This condition is called complete ionization. Under a complete ionization condition, we can write the electron density as

$$n = N_D$$

where N_D is the donor concentration. We obtain the Fermi

level in term of the effective density of states N_C and the donor concentration N_D

$$E_C - E_F = kT \ln[N_C / N_D]$$

From equation it is clear that the higher the donor concentration, the smaller the energy difference ($E_C, - E_F$), that is, the Fermi level will move closer to the bottom of the conduction band. Similarly, for higher acceptor concentrations, the Fermi level will move closer to the top of the valence band. Figure 1 – 6 shows the graphic procedure to obtain the carrier concentration. This is similar to that shown in Figure 1 – 5. However, the Fermi level is closer to the bottom of the conduction band, and the electron concentration (upper shaded area) is much larger than the hole concentration (lower shaded area). Because of the mass action law, the product of n and p is the same as that for the intrinsic case (i.e., $np = n_i^2$).

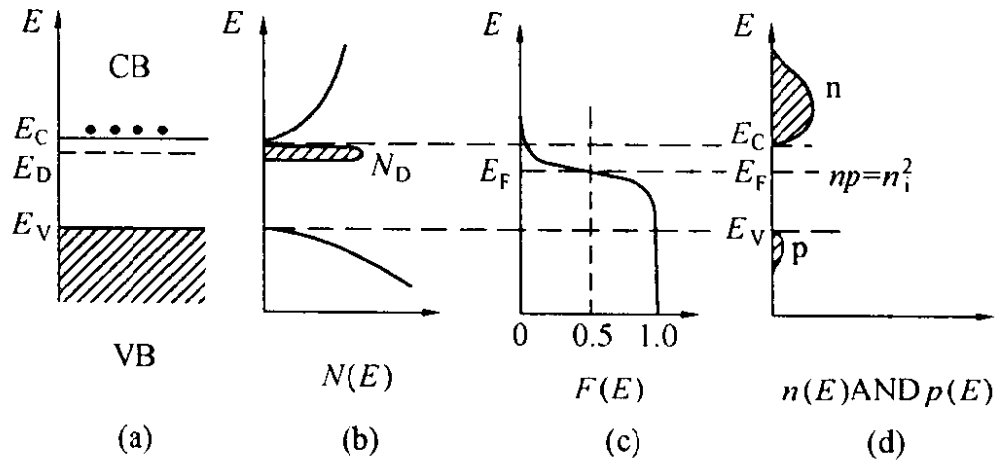


Figure 1 – 6 n-type semiconductor

1.2 Carrier Transport Phenomena

In this section we consider various transport phenomena that arise from the motion of charge carriers (electrons and holes) in semiconductors under the influence of an electric field and a

carrier concentration gradient. We shall discuss the concept of injection of excess carriers, which gives rise to a none-equilibrium condition, that is, the carrier concentration product pn is different from its equilibrium value n_i^2 . Return to an equilibrium condition through the generation-recombination processes will be considered next. We shall then derive the basic governing equations for semiconductor device operation, which include the current density equation and the continuity equation. The section closes with a brief discussion of high-field effects, which result in velocity saturation and impact ionization.

1.2.1 Carrier Drift

Consider an *n-type* semiconductor sample with uniform donor concentration in thermal equilibrium. As discussed in Section 1.1, the conduction electrons in the semiconductor conduction band are essentially free particles, since they are not associated with any particular lattice or donor site. The influence of crystal lattices is incorporated, in the effective mass of conduction electrons, which differs somewhat from the mass of free electrons. Under thermal equilibrium, the average thermal energy of a conduction electron can be obtained from the theorem for equipartition of energy: $1/2 kT$ units of energy per degree of freedom, where k is Boltzmann's constant and T is the absolute temperature. The electrons in a semiconductor have three degrees of freedom; they can move about in a three-dimensional space. Therefore, the kinetic energy of the electrons is given by

$$\frac{1}{2} m_n v_{th}^2 = \frac{3}{2} kT \quad (1-13)$$

where m_n is the effective mass of electrons and v_{th} is the

average thermal velocity. At room temperature (300 K) the thermal velocity is about 10^7 cm/s for silicon and gallium arsenide.

The electrons in the semiconductor are therefore moving rapidly in all directions. The thermal motion of an individual electron may be visualized as a succession of random scattering from collisions with lattice atoms, impurity atoms, and other scattering centers, as illustrated in Figure 1 – 7(a). The random motion of electrons leads to a zero net displacement of an electron over a sufficiently long period of time. The average distance between collisions is called the mean free path, and the average time between collisions is called the mean free time τ_c . For a typical value of 10^{-5} cm for the mean free path, τ_c is about 1 ps (i. e., $10^{-5}/v_{th} \approx 10^{-12}$ s).

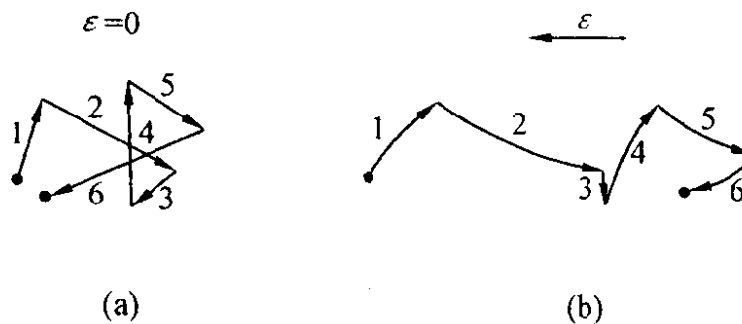


Figure 1 – 7 Schematic path of an electron in a semiconductor

When a small electric field E is applied to the semiconductor sample, each electron will experience a force $-qE$ from the field and will be accelerated along the field (in the opposite direction to the field) during the time between collisions. Therefore, an additional velocity component will be superimposed upon the thermal motion of electrons. This additional component is called the drift velocity. The combined displacement of an electron due to the random thermal motion and the drift component is illustrated in Figure 1 – 7(b). Note that there is a net displacement

of the electron in the direction opposite to the applied field.

We can obtain the drift velocity v , by equating the momentum (force $\times$ time) applied to an electron during the free flight between collisions to the momentum gained by the electron in the same period. The equality is valid because in a steady state all momentum gained between collisions is lost to the lattice in the collision. The momentum applied to an electron is given by $-qE$, and the momentum gained is $m_n v_n$. We have

$$-q\epsilon\tau_c = m_n v_n \quad (1 - 14)$$

or

$$v_n = - \left[\frac{q\tau_c}{m_n} \right] \epsilon \quad (1 - 14a)$$

Equation 1 - 14a states that the electron drift velocity is proportional to the applied electric field. The proportionality factor depends on the mean free time and the effective mass. The proportionality factor is called the electron mobility μ_n , in units of cm^2/Vs , or

$$\mu_n \equiv \frac{q\tau_c}{m_n} \quad (1 - 15)$$

Thus

$$v_n = - \mu_n \epsilon \quad (1 - 16)$$

Mobility is an important parameter for carrier transport because it describes how strongly the motion of an electron is influenced by an applied electric field. A similar expression can be written for holes in the valence band

$$v_p = \mu_p \epsilon \quad (1 - 17)$$

where v_p is the hole drift velocity and μ_p is the hole mobility. The negative sign is removed in Eq. 1 - 17 because holes drift in the same direction as the electric field.

In Eq. 1 – 15 the mobility is related directly to the mean free time between collisions, which in turn is determined by the various scattering mechanisms. The two most important mechanisms are lattice scattering and impurity scattering. Lattice scattering results from thermal vibrations of the lattice atoms at any temperature above absolute zero. These vibrations disturb lattice periodic potential and allow energy to be transferred between the carriers and the lattice. Since lattice vibration increases with increasing temperature, lattice scattering becomes dominant at high temperatures; hence the mobility decreases with increasing temperature. Theoretical analysis shows that the mobility due to lattice scattering μ_L will decrease in proportion to $T^{-3/2}$.

Impurity scattering results when a charge carrier travels past an ionized dopant impurity (donor or acceptor). The charge carrier path will be deflected due to Coulomb force interaction. The probability of impurity scattering depends on the total concentration of ionized impurities, that is, the sum of the concentration of negatively and positively charged ions. However, unlike lattice scattering, impurity scattering becomes less significant at higher temperatures. At higher temperatures, the carriers move faster; they remain near the impurity atom for a shorter time and are therefore less effectively scattered. The mobility due to impurity scattering μ_I can theoretically be shown to vary as $T^{3/2}/N_T$, where N_T is the total impurity concentration.

The probability of a collision taking place in unit time, $1/\tau_c$, is the sum of the probabilities of collisions due to the various scattering mechanisms.

The measured electron mobility as a function of temperature

for silicon with five different donor concentrations is shown. The insert shows the theoretical temperature dependence of mobility due to both lattice and impurity scatterings. For lightly doped samples (e.g., the sample with doping of 10^{14} cm^{-3}), the lattice scattering dominates, and the mobility decreases as the temperature increases. For heavily doped samples, the effect of impurity scattering is most pronounced at low temperatures. The mobility increases as the temperature increases, as can be seen for the sample with doping of 10^{19} cm^{-3} . For a given temperature, the mobility decreases with increasing impurity concentration, because of enhanced impurity scatterings.

The measured mobilities in silicon and gallium arsenide as a function of impurity concentration at room temperature are shown. Mobility reaches a maximum value at low impurity concentrations; this corresponds to the lattice-scattering limitation. Both electron and hole mobilities decrease with increasing impurity concentration and eventually approach a minimum value at high concentrations. Note also that the mobility of electrons is greater than that of holes. Greater electron mobility is due mainly to the smaller effective mass of electrons.

1.2.2 Carrier Diffusion

In the preceding section we considered the drift current, that is, the transport of carriers when an electric field is applied. Another important current component can exist if there is a spatial variation of carrier concentration in the semiconductor material, that is, the carriers tend to move from a region of high concentration to a region of low concentration. This current component is called diffusion current.

To understand the diffusion process, let us assume an electron density that varies in the x -direction. The semiconductor is at uniform temperature, so that the average thermal energy of electrons does not vary with x , only the density $n(x)$ varies. We shall consider the number of electrons crossing the plane at $x = 0$ per unit time and per unit area. Because of finite temperature, the electrons have random thermal motions with a thermal velocity v_{th} and a mean free path l . (Note that $l = v_{th}\tau_c$, where τ_c is the mean free time.) The electrons at $x = -l$, one mean free path away on the left side, have equal chances of moving left or right; and in a mean free time τ_c , one half of time will move across the plane $x = 0$.

Because each electron carries a charge- q , the carrier flow gives rise to a current

$$J_n = qD_n \frac{dn}{dx} \quad (1-18)$$

Where $D_n \equiv v_{th}l$ is called diffusivity.

The diffusion current is proportional to the spatial derivative of the electron density. Diffusion current results from the random thermal motion of carriers in a concentration gradient. For an electron density that increases with x , the gradient is positive, and the electrons will diffuse toward the negative x -direction. The current is positive and flows in the direction opposite to that of the electrons

1.2.3 Carrier Injection

In thermal equilibrium the relationship $pn = n_i^2$ is valid. If excess carriers are introduced to a semiconductor so that $pn > n_i^2$, we have a none-equilibrium situation. The process of introducing

excess carriers is called carrier injection. We can inject carriers by using various methods including optical excitation and forward biasing a p-n junction (discussed in Section 1.3). In the case of optical excitation, we shine a light on a semiconductor. If the photon energy $h\nu$ of the light is greater than the bandgap energy E_g of the semiconductor, where h is the Planck constant and ν is the optical frequency, the photon is absorbed by semiconductor and an electron-hole pair is generated. The optical excitation increases the electron and hole concentrations above their equilibrium values. These additional carriers are called *excess carriers*.

The magnitude of the excess carrier concentration relative to the majority carrier concentration determines the injection level. We shall use an example to clarify the meaning of injection level. The majority carrier concentration is approximately equal to the donor concentration, that is, $n_{no} = 10^5 \text{ cm}^{-3}$. The minority carrier concentration is given by $p_{no} = n_i^2/n_{no} = 1.45 \times 10^5 \text{ cm}^{-3}$. In this notation, the first subscript refers to the type of semiconductor and the subscript o refers to the thermal equilibrium condition. Thus, n_{no} and p_{no} denote the electron and hole concentrations, respectively, in an n -type semiconductor in equilibrium.

When we introduce (e. g., by optical excitation) excess carriers of both types into the semiconductor, the excess electron concentration Δn must equal the excess hole concentration Δp , because electrons and holes are produced in pairs. In the example shown in Figure 1 - 8 (b), we have added $10^{12}/\text{cm}^2$ minority carriers (holes in an n -type semiconductor). Therefore, the hole concentration is increased by seven orders of magnitude (from $10^5/\text{cm}^3$ to $10^{12}/\text{cm}^3$). At the same time we have added 10^{12} cm^{-3} majority carriers (electrons) to the semiconductor. However, this

concentration of excess electrons is negligibly small compared to the original electron concentration. The percentage change in the majority carrier concentration is only 0.1%. This condition, in which the excess carrier concentration is small in comparison to the doping concentration, that is, $\Delta n = \Delta p \ll N_D$, is referred to as low-level injection.

Figure 1 – 8 shows an example of high-level injection in which the number of injected excess carriers is comparable to or larger than the number of carriers due to doping concentration. In this case, the injected-carrier concentrations may overwhelm the equilibrium majority carrier concentration, and p , becomes comparable to n , as indicated in the figure. High-level injection is sometime encountered in device operations. However, because of the complexities involved in its treatment, we shall be concerned mainly with low-level injection.

1.2.4 Generation and Recombination Processes

Whenever the thermal-equilibrium condition is disturbed (i. e. , $pn \neq n_i^2$), processes exist to restore the system to equilibrium (i. e. , $pn = n_i^2$). In the case of the injection of excess carriers, the mechanism that restores equilibrium is recombination of the injected minority carriers with the majority carriers. Depending on the nature of the recombination process, the released energy that results from the recombination process can be emitted as a photon or dissipated as heat to the lattice. When a photon is emitted, the process is called radiative recombination; otherwise, it is called nonradiative recombination.

Recombination phenomena can be classified as direct and indirect processes. Direct recombination, also called band-to-band

recombination, usually dominates in direct-bandgap semiconductors, such as gallium arsenide; while indirect recombination via bandgap recombination centers dominates in indirect bandgap semiconductors, such as silicon.

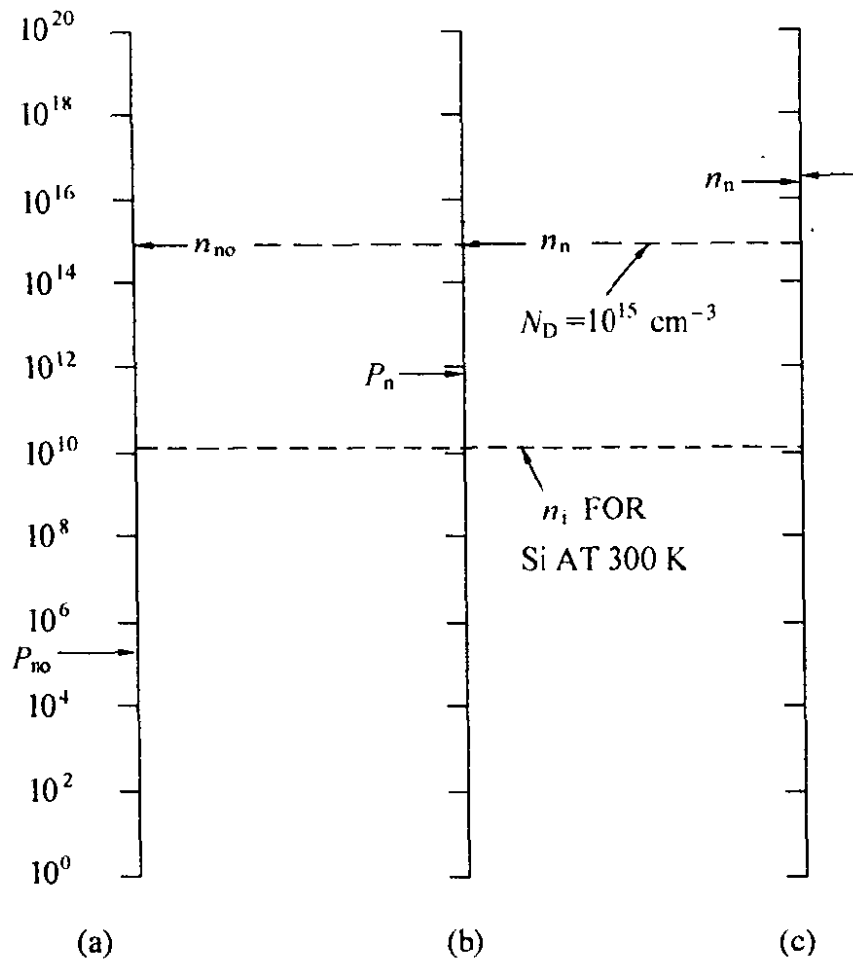


Figure 1 - 8 Carrier concentration in n-type Si

1) Direct Recombination

Consider a direct-band gap semiconductor in thermal equilibrium. The continuous thermal vibration of lattice atoms causes some bonds between neighboring atoms to be broken. When a bond is broken, an electron-hole pair is generated. In terms of the band diagram, the thermal energy enables a valence electron to make an upward transition to the conduction band leaving a hole in the valence band. This process is called carrier

generation and is represented by the generation rate G_{th} (number of electron-hole pairs generated per cm^3 per second) in Figure 1 - 9(a). When an electron makes a transition downward from the conduction band to the valence band, an electron-hole pair is annihilated. This reverse process is called recombination; it is represented by the recombination rate R_{th} in Figure 1 - 9(a). Under thermal equilibrium conditions, the generation rate G_{th} must equal the recombination rate R_{th} , so that the carrier concentrations remain constant and the condition $pn = n_i^2$ is maintained.

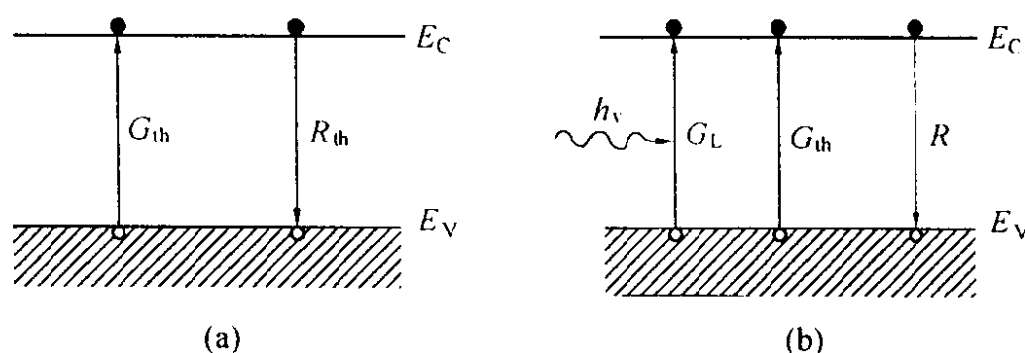


Figure 1 - 9 Direct generation and recombination for electron-hole pairs

When excess carriers are introduced to a direct-bandgap semiconductor, the probability is high that electrons and holes will recombine directly, because the bottom of the conduction band and the top of the valence band are lined up and no additional crystal momentum is required for the transition across the bandgap. The rate of the direct recombination R is expected to be proportional to the number of holes available in the valence band; that is

$$R = \beta np$$

Where β is the proportionality constant. As discussed previously, in thermal equilibrium the recombination rate must be

balanced by the generation rate. Therefore, for an n-type semiconductor, we have

$$G_{th} = R_{th} = \beta n_{n0} p_{n0}$$

Where n_{n0} and p_{n0} represent electron and hole densities in an n-type semiconductor at thermal equilibrium. When we shine a light on the semiconductor to produce electron-hole pairs at a rate G_L [Figure1 – 9(b)], the carrier concentrations are above their equilibrium values.

Therefore, the net recombination rate is proportional to the excess minority carrier concentration. Obviously, $U = 0$ in thermal equilibrium. The proportionality constant $1/\beta n_{n0}$ is called the lifetime τ_p of the excess minority carriers. The physical meaning of lifetime can best be illustrated by the transient response of a device after the sudden removal of the light source. Consider an n-type sample, as shown in Figure1 – 10(a), that is illuminated with light and in which the electron-hole pairs are generated uniformly throughout the sample with a generation rate G_L . Figure1 – 10(b) shows the variation of pn with time. The minority carriers recombine with majority carriers and decay exponentially with a time constant τ_p which corresponds to the lifetime.

The above case illustrates the main idea of measuring the carrier lifetime using photoconductivity method. Figure1 – 10(c) shows a schematic setup. The excess carriers, generated uniformly throughout the sample by the light pulse, cause a momentary increase in the conductivity. The increase in conductivity manifests itself by a drop in voltage across the sample when a constant current is passed through it. The decay of the conductivity can be observed on an oscilloscope and is a measure of the lifetime of the excess minority carriers.

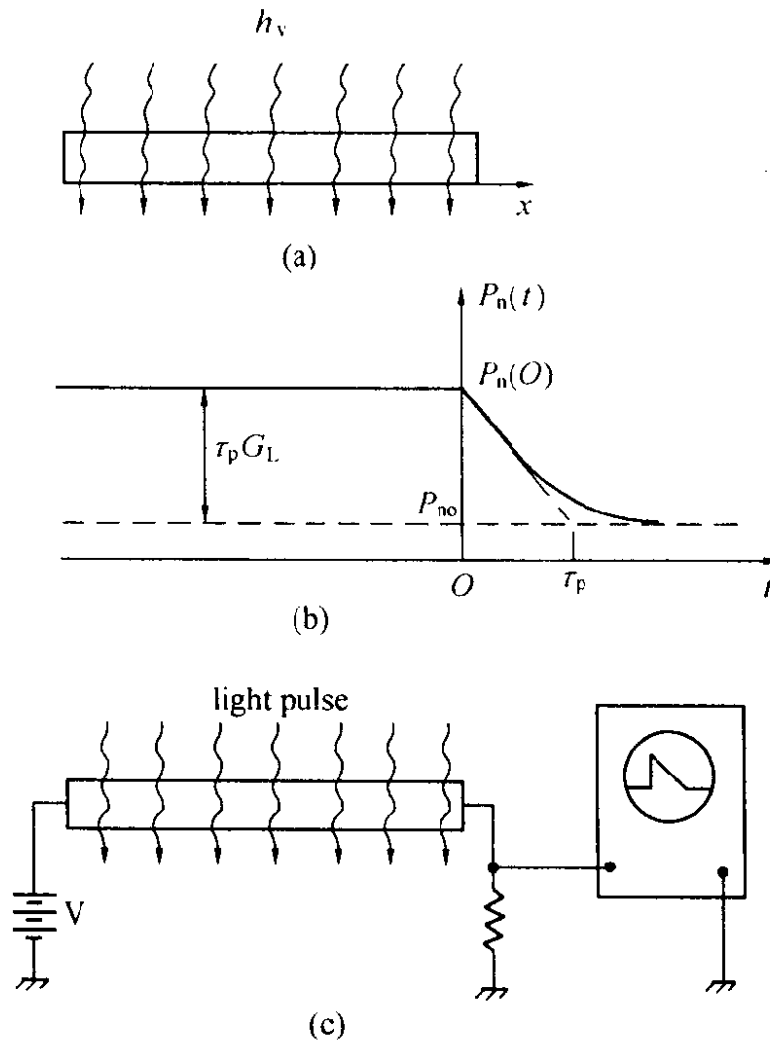


Figure 1 - 10 Decay of photoexcited carriers

1.3 PN Junction

Most semiconductor devices contain at least one junction between p-type and n-type material. These p-n junctions are fundamental to the performance of functions such as rectification, amplification, switching, and other operations in electronic circuits. In this section we shall discuss the equilibrium state of junction and the flow of electrons and holes across a junction under steady state and transient conditions.

1.3.1 Equilibrium Conditions

In this section we wish to develop both a useful mathematical description of the p-n junction and a strong qualitative understanding of its properties. There must be some compromise in these two goals, since a complete mathematical treatment would obscure the essentially simple physical features of junction operation; on the other hand, a completely qualitative description would not be useful in making calculations. The approach, therefore, will be to describe the junction mathematically while neglecting small effects, which add little to the basic solution.

The mathematics of p-n junctions is greatly simplified for the case of the step junction, which has uniform p doping on one side of a sharp junction and uniform n doping on the other side. This model represents the alloyed junction quite well; the diffused junction, however, is actually graded ($N_d - N_a$ varies over a significant distance on either side of the junction). After the basic ideas of junction theory are explored for the step junction, we can make the appropriate corrections to extend the theory to the graded junction. In these discussions we shall assume one-dimensional current flow in samples of uniform cross-sectional area.

In this section we investigate the properties of the step junction at equilibrium (i. e. , with no external excitation and no net currents flowing in the device). We shall find that the difference in doping on each side of the junction causes a potential difference between the two types of material. This is a reasonable result, since we would expect some charge transfer because of diffusion between the p material (many holes) and the n material

(many electrons). In addition, we shall find that there are four components of current, which flow across the junction due to the drift and diffusion of electrons and hole. These four components combine to give zero net current for the equilibrium case. However, the application of bias to the junction increases some of these current components with respect to others, giving net current flow. If we understand the nature of these four current components, a sound view of p-n junction operation, with or without bias, will follow.

Let us consider separate regions of p and n-type semiconductor material, brought together to form a junction (Figure 1 - 11). This is not a practical way of forming a device, but it does allow us to discover the requirements of equilibrium at a junction. Before they are joined, the n material has a large concentration of electrons and few holes, whereas the converse is true for the p material. Upon joining the two regions, we expect diffusion of carriers to take place because of the large carrier concentration gradients at the junction. Thus holes diffuse from the p side into the n side, and electrons diffuse from n to p. The resulting diffusion current cannot build up indefinitely, however, because an opposing electric field is created at the junction. If the two regions were boxes of red air molecules and green molecules (perhaps due to appropriate types of pollution), eventually there would be a homogeneous mixture of the two after the boxes were joined. This cannot occur in the case of the charged particles in a p-n junction because of the development of space charge and the electric field. If we consider that electrons diffusing from n to p leave behind uncompensated donor ions (N_d^+) in the n material, and holes leaving the p region leave behind uncompensated

acceptors (Na^-), it is easy to visualize the development of a region of positive space charge near the n side of the junction and negative charge near the p side. The resulting electric field is directed from the positive charge toward the negative charge. Thus ϵ is in the direction opposite to that of diffusion current for each type of carrier (recall electron current is opposite to the direction of electron flow). Therefore, the field creates a drift component of current from n to p, opposing the diffusion current.

Since we know that no net current can flow across the junction at equilibrium, the current due to the drift of carriers in the E field must exactly cancel the diffusion current. Furthermore, since there can be no net buildup of electrons or holes on either side as a function of time, the drift and diffusion currents must cancel for each type of carrier

$$J_p(\text{drift}) + J_p(\text{diff.}) = 0$$

$$J_n(\text{drift}) + J_n(\text{diff.}) = 0$$

Therefore, the electric field (ϵ) builds up to the point where the net current is zero at equilibrium. The electric field appears in some region W about the junction, and there is an equilibrium potential difference V_0 across W . In the electrostatic potential diagram of Figure 1-11(b), there is a gradient in potential in the direction opposite to ϵ , in accordance with the fundamental relation. $\epsilon(x) = dv(x)/dx$. We assume the electric field is zero in the neutral regions outside W . Thus there is a constant potential v_n in the neutral n material, a constant v_p in the neutral p material, and a potential difference $V_0 = v_n - v_p$ between the two. The region W is called the transition region, and the potential difference V_0 is called the contact potential. The contact potential appearing across W is a built-in potential barrier, in that it is

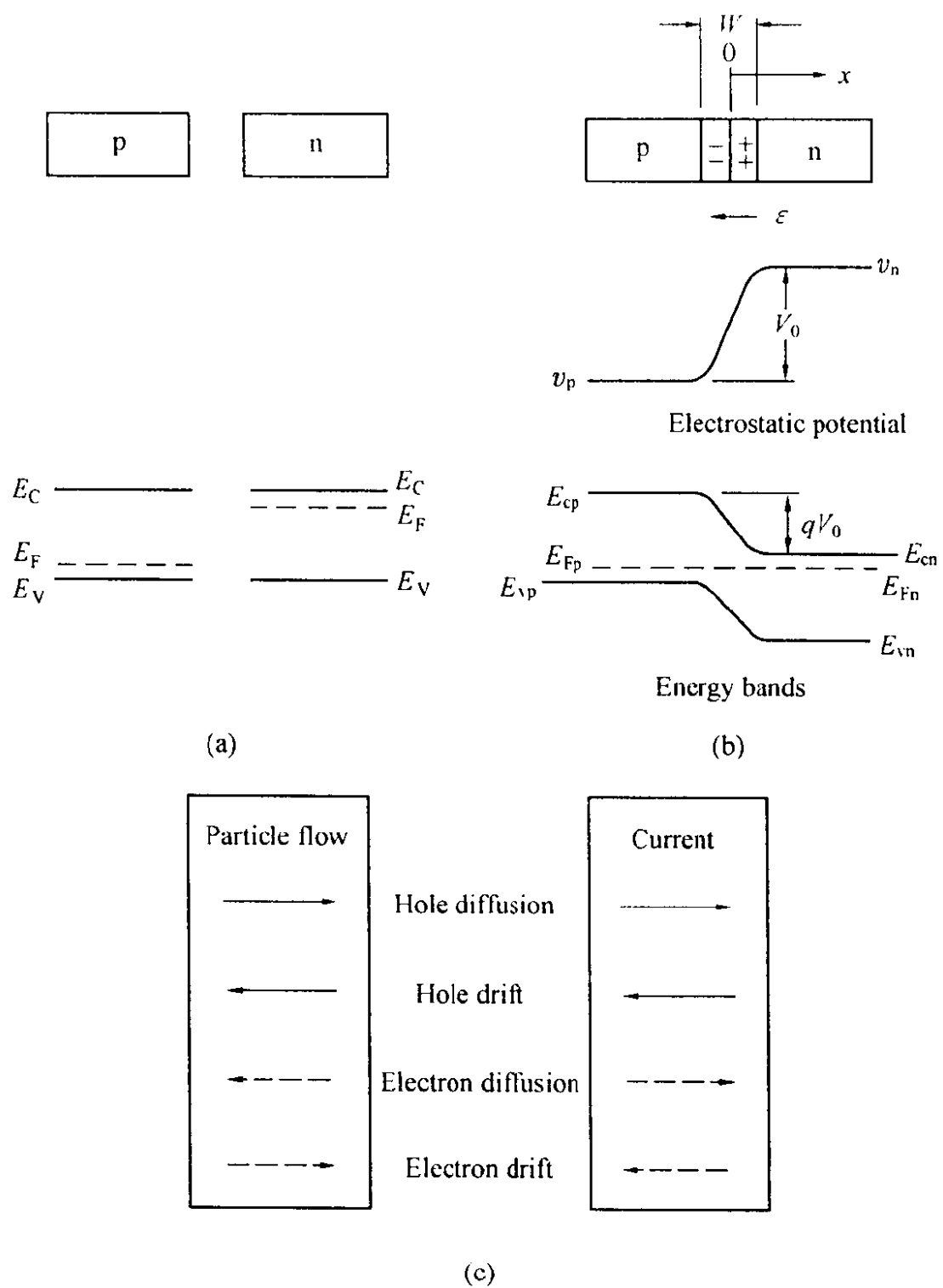


Figure 1 - 11 Properties of an equilibrium p-n junction

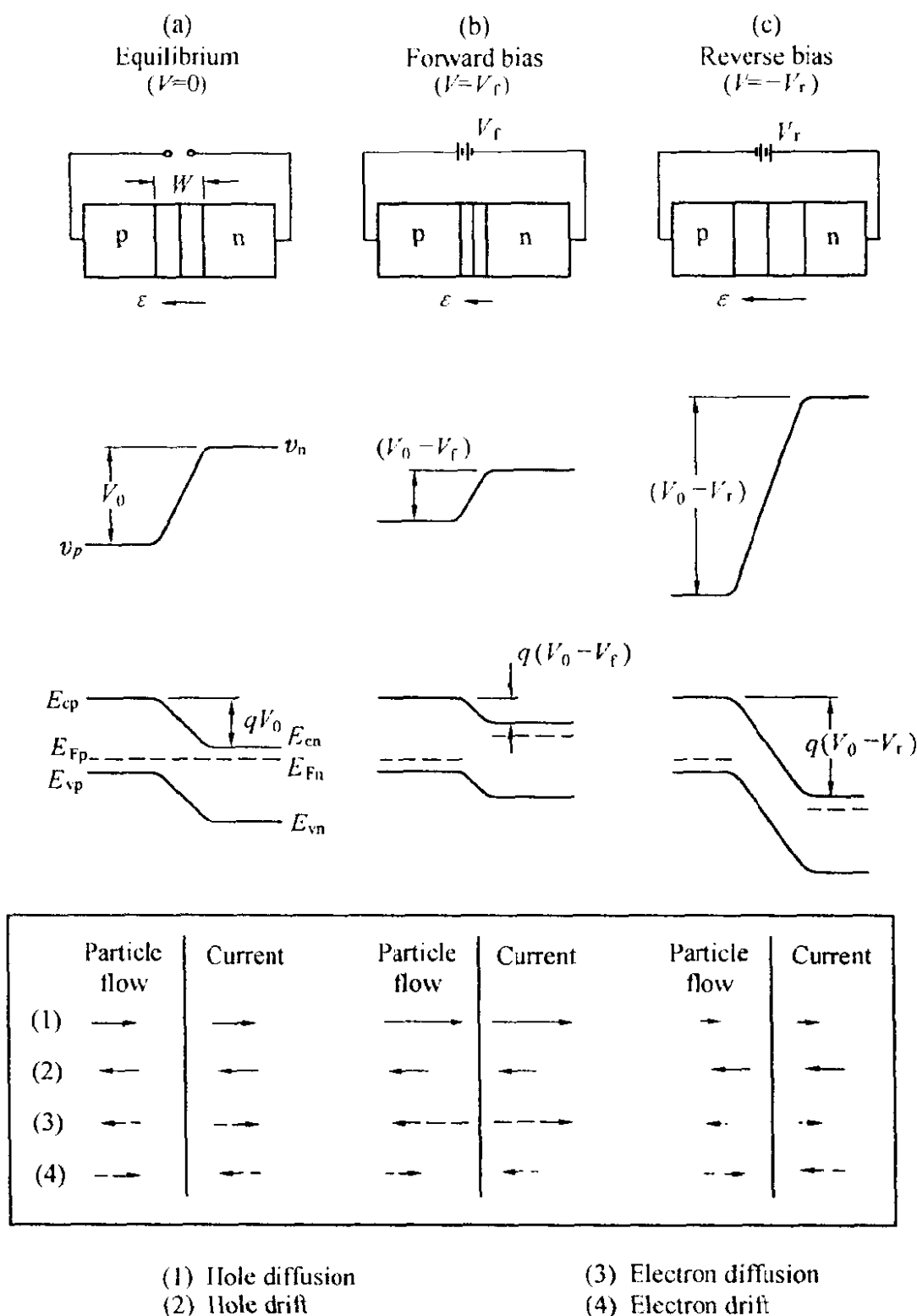


Figure 1 - 12 Effects of bias at a p-n junction; transition region width and electric field, electrostatic potential, energy band diagram, and particle flow and current directions within W for (a) equilibrium, (b) forward bias, and (c) reverse bias

necessary to the maintenance of equilibrium at the junction; it does not imply any external potential. Indeed, the contact potential cannot be measured by placing a voltmeter across the device, because new contact potentials are formed at each probe, just canceling V_0 . By definition V_0 is an equilibrium quantity, and no net current can result from it.

The contact potential separates the bands as in Figure 1 – 11 (b); the valence and conduction energy bands are higher on the p side of the junction than on the n side by the amount qV_0 . The separation of the bands at equilibrium is just that required to make the Fermi level constant throughout the device. We shall show that this is the case in the following section. However, we could predict the lack of spatial variation of the Fermi level from thermodynamic arguments. Any gradient in the quasi-Fermi level implies a net current. Since $E_F = F_n = F_p$ at equilibrium, and since the net current must be zero, we conclude that EF must be constant across the junction. Thus if we know the band separation for the junction simply by drawing a diagram such as Figure 1 – 11 (b) with the Fermi levels aligned.

To obtain a quantitative relationship between V_0 and the doping concentrations on each side of the junction, we must use the requirements for equilibrium in the drift and diffusion current equations. The drift and diffusion components of the hole current just cancel at equilibrium

$$J_p(x) = q[\mu_p p(x) \epsilon(x) - D_p \frac{dp(x)}{dx}] = 0 \quad (1 - 19)$$

1.3.2 Forward-And Reverse-Biased Junctions

One useful feature of a p-n junction is that current flows quite

freely in the p to n direction when the p region has a positive external voltage bias relative to n (forward bias and forward current), whereas virtually no current flows when p is made negative relative to n (reverse bias and reverse current). This asymmetry of the current flow makes the p-n junction diode very useful as a rectifier. As an example of rectification, suppose a sinusoidal a-c generator is placed in series with a resistor and a diode, which passes current only in one direction. The resulting current through the resistor will reflect only half of the sinusoidal signal—for example, only the positive half-cycles. The rectified sine wave has an average value and can be used, for example, to charge a storage battery; on the other hand, the input sine wave has no average value. Diode rectifiers are useful in many applications in electronic circuits, particularly in “wave-shaping” (making use of the nonlinear nature of the diode to change the shape of time-varying signals).

While rectification is an important application, it is only the beginning of a host of uses for the biased junction. Biased p-n junctions can be used as voltage-variable capacitors, photocells, light emitters, and many more devices, which are basic to modern electronics. Two or more junctions can be used to form transistors and controlled switches.

In this section we begin with a qualitative description of current flow in a biased junction. With the background of the previous section, the basic features of current flow are relatively simple to understand, and these qualitative concepts form the basis for the analytical description of forward and reverse currents in a junction.

We assume that an applied voltage bias V appears across the

transition region of the junction rather than in the neutral n and p regions. Of course, there will be some voltage drop in the neutral material, if a current flows through it. But in most p-n junction devices, the length of each region is small compared with its area, and the doping is usually moderate to heavy; thus the resistance is small in each neutral region, and only a small voltage drop can be maintained outside the space charge (transition) region. For almost all calculations it is valid to assume that an applied voltage appears entirely across the transition region. We shall take V to be positive when the external bias is positive on the p side relative to the n side.

Since an applied voltage changes the electrostatic potential barrier and thus the electric field within the transition region, we would expect changes in the various components of current at the junction (Figure 1 - 12). In addition, the separation of the energy bands is affected by the applied bias, along with the width of the depletion region. Let us begin by examining qualitatively the effects of bias on the important features of the junction.

The electrostatic potential barrier at the junction is lowered by a forward bias V_f from the equilibrium contact potential V_0 to the smaller value $V_0 - V_f$. This lowering of the potential barrier occurs because a forward bias (p positive with respect to n) raises the electrostatic potential on the p side relative to the n side. For a reverse bias ($V = -V_r$) the opposite occurs; the electrostatic potential of the p side is depressed relative to the n side, and the potential barrier at the junction becomes larger ($V_0 + V_r$).

The electric field within the transition region can be deduced from the potential barrier. We notice that the field decreases with forward bias, since the applied electric field opposes the built-in

field. With reverse bias the field at the junction is increased by the applied field, which is in the same direction as the equilibrium field.

The change in electric field at the junction calls for a change in the transition width W , since it is still necessary that a proper number of positive and negative charges (in the form of uncompensated donor and acceptor ions) be exposed for a given value of the field. Thus we would expect the width W to decrease under forward bias (smaller ϵ , fewer uncompensated charges) and to increase under reverse bias.

The separation of the energy bands is a direct function of the electrostatic potential barrier at the junction. The height of the electron energy barrier is simply the electronic charge q times the height of the electrostatic potential barrier. Thus the bands are separated less $q (V_0 - V_f)$ under forward bias than at equilibrium, and more under reverse bias $q (V_0 - V_r)$. We assume the Fermi level deep inside each neutral region is essentially the equilibrium value (we shall return to this assumption later); therefore, the shifting of the energy bands under bias implies a separation of the Fermi levels on either side of the junction. Under forward bias, the Fermi level on the n side E_{Fn} is above E_{Fp} by the energy qV_f ; for reverse bias, E_{Fp} is qV_r joules higher than E_{Fn} . In energy units of electron volts, the Fermi levels in the two neutral regions are separated by applied voltage.

The diffusion current is composed of majority carrier electrons on the n side surmounting the potential energy barrier to diffuse to the p side, and holes surmounting their barrier from p to n. There is a distribution of energies for electrons in the n-side conduction band, and some electrons in the high-energy "tail" of the

distribution have enough energy to diffuse from n to p at equilibrium in spite of the barrier. With forward bias, however, the barrier is lowered (to $V_0 - V_f$), and many more electrons in the n-side conduction band have sufficient energy to diffuse from n to p quite large with forward bias. Similarly, more holes can diffuse from p to n under forward bias because of the lowered barrier. For reverse bias the barrier becomes so large ($V_0 + V_f$) that virtually no electrons in the n-side conduction band or holes in the p-side valence band have enough energy to surmount it. Therefore, the diffusion current is usually negligible for reverse bias.

The drift current is relatively insensitive to the height of the potential barrier. This sounds strange at first, since we normally think in terms of material with ample carriers, and therefore we expect drift current to be simply proportional to the applied field. The reason for this apparent anomaly is the fact that the drift current is limited not by how fast carriers are swept down the barrier, but rather how often. For example, minority carrier electrons on the p side, which wander into the transition region will be swept down the barrier by the field, giving rise to the electron component of drift current. However, this current is small not because of the size of the barrier, but because there are very few minority electrons in the p side to participate. Every electron on the p side, which diffuses to the transition region will be swept down the potential energy hill, whether the hill is large or small. The electron drift current does not depend on how fast a individual electron is swept from p to n, but rather on how many electrons are swept down the barrier per second. Similar comments apply regarding the drift of minority holes from the n

side to the p side of the junction. To a good approximation, therefore, the electron and hole drift currents at the junction are independent of the applied voltage.

The supply of minority carriers on each side of the junction required to participate in the drift component of current is generated by thermal excitation of electron-hole pairs. For example, an EHP created near the junction on the p side provides a minority electron in the p material. If the EHP is generated within a diffusion length L_n of the transition region, this electron can diffuse to the junction and be swept down the barrier to the n side. The resulting current due to drift of generated carriers across the junction is commonly called the generation current since its magnitude depends entirely on the rate of generation of EHPs. As we shall discuss later, this generation current can be increased greatly by optical excitation of EHPs near the junction (the p-n junction photodiode).

The total current crossing the junction is composed of the sum of the diffusion and drift components. As Figure 1 - 12 indicates, the electron and hole diffusion currents are both directed from p to n (although the particle flow directions are opposite to each other), and the drift currents are from n to p. the net current crossing the junction is zero at equilibrium, since the drift and diffusion components cancel for each type of carrier (the equilibrium electron and hole components need not be equal, as in Figure 1 - 12, as long as the net hole current and the net electron current are each zero). Under reverse bias, both diffusion components are negligible because of the large barrier at the junction, and the only current is the relatively small (and essentially voltage-independent) generation current from n to p.

This generation current is shown in Figure 1 – 13, in a sketch of a typical $I - V$ plot for a p-n junction. In this figure the positive direction for the current I is taken from p to n, and the applied voltage V is positive when the positive battery terminal is connected to p and the negative terminal to n. The only current flowing in this p-n junction diode for negative V is the small current $I(\text{gen.})$ due to carriers generated in the transition region or minority carriers which diffuse to the junction and are collected. The current at $V = 0$ (equilibrium) is zero since the generation and diffusion currents cancel

$$I = I(\text{diff.}) - |I(\text{gen.})| = 0$$

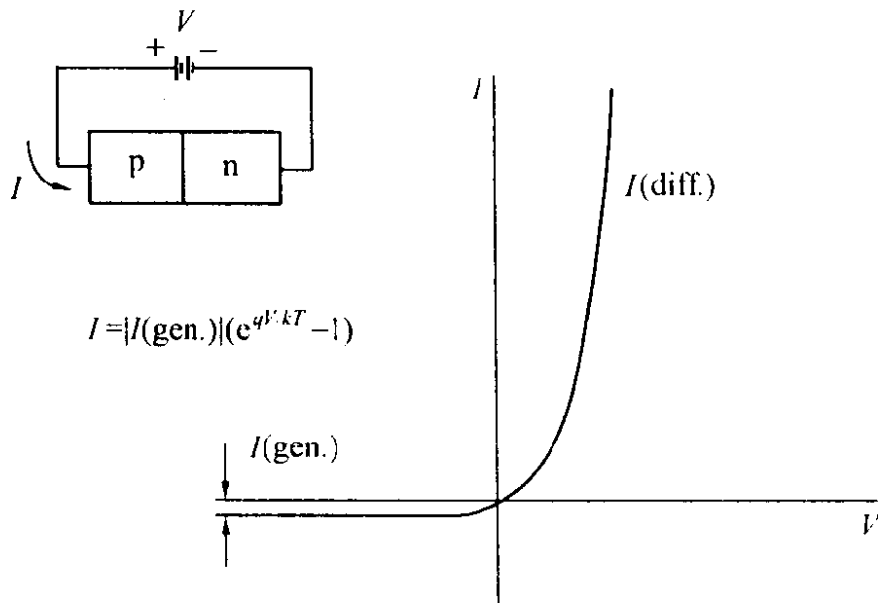


Figure 1 – 13 I-V characteristic of a p-n junction

As we shall see in the next section, an applied forward bias $V = V_f$ increases the probability that a carrier can diffuse across the junction, by the factor $\exp (qV_f/kT)$. Thus the diffusion current under forward bias is given by its equilibrium value multiplied by $\exp (qV/kT)$; similarly, for reverse bias the diffusion current is the equilibrium value reduced by the same

factor, with $V = -V_r$. Since the equilibrium diffusion current is equal in magnitude to $|I(\text{gen.})|$, the diffusion current with applied bias is simply $|I(\text{gen.})| \exp(qV/kT)$. The total current I is then the diffusion current minus the absolute value of the generation current, which we will now refer to as I_0

$$I = I_0(e^{qV/kT} - 1) \quad (1 - 20)$$

In Eq. (1 - 20) the applied voltage V can be positive or negative, $V = V_f$ or $V = -V_r$. When V is positive and greater than a few kT/q ($kT/q = 0.0259$ V at room temperature), the exponential term is much greater than unity. The current thus increases exponentially with forward bias. When V is negative (reverse bias), the exponential term approaches zero and the current is $-I_0$, which is in the n to p (negative) direction. This negative generation current is also called the *reverse saturation current*. The striking feature of Figure 1 - 13 is the nonlinearity of the I - V characteristic. Current flows relatively freely in the forward direction of the diode, but almost no current flows in the reverse direction.

From the discussion in the previous section we expect the minority carrier concentration on each side of a p-n junction to vary with the applied bias because of variations in the diffusion of carriers across the junction. The equilibrium ratio of hole concentrations on each side

$$p_p/p_n = e^{qV_0/kT} \quad (1 - 21)$$

Becomes with bias

$$p(-x_{p0})/p(x_{n0}) = e^{q(V_0 - V)/kT} \quad (1 - 22)$$

This equation is obtained by the same mathematics as Eq. (1 - 21), but with the altered barrier $V_0 - V$; it relates the steady state hole concentrations on the two sides of the transition

region with either forward or reverse bias (V positive or negative). For low-level injection we can neglect changes in the majority carrier concentrations, which vary only slightly with bias compared with their equilibrium values. With this simplification we can write the ratio of Eq. (1 - 21) to (1 - 22) as

$$p(x_{no})/p_n = e^{qV/kT} \quad (1 - 23)$$

With forward bias, Eq. (1 - 23) suggests a greatly increased minority carrier hole concentration at the edge of the transition region on the n side $p(x_{no})$ than was the case at equilibrium. Conversely the hole concentration $p(x_{no})$ under reverse bias (V negative) is reduced below the equilibrium value p_n . The exponential increase of the hole concentration at x_{no} with forward bias is an example of minority carrier injection. We can easily calculate the excess hole concentration Δp_n at the edge of the transition region by subtracting the equilibrium hole concentration from Eq. (1 - 23)

$$\Delta p_n = p(x_{no}) - p_n = p_n(e^{qV/kT} - 1) \quad (1 - 24)$$

From our study of diffusion of excess carriers in last section, we expect that injection leading to a steady concentration of Δp_n excess holes at x_{no} will produce a distribution of excess holes in the n material. As the holes diffuse deeper into the n region, they recombine with electrons in the n material, and the resulting excess hole distribution is obtained as a solution of the diffusion equation. If the n region is long compared with the hole diffusion length L_p the solution is exponential. Similarly, the injected electrons in the p material diffuse and recombine, giving an exponential distribution of excess electrons. For convenience, let us define two new coordinates: distances measured in the x -direction in the n material from x_{no} will be designated x_n ;

distances in the p material measured in the $-x$ -direction with $-x_{po}$ as the origin will be called x_{po} . This convention will simplify the mathematics considerably. We can write the diffusion equation for each side of the junction and solve for the distributions of excess carriers assuming long n and p regions

$$\delta p(x_n) = \Delta p_n e^{-x_p/L_n} = p_n(e^{qV/kT} - 1) e^{-x_p/L_p} \quad (1-25a)$$

$$\delta n(x_n) = \Delta p n_{np} e^{-x_n/L_p} = n_p(e^{qV/kT} - 1) e^{-x_n/L_n} \quad (1-25b)$$

The hole diffusion current at any point x_n in the n material can be calculated

$$I_p(x_n) = -qAD_p d\delta_p(x_n)/dx_n = qAD_p/L_p \delta p(x_n) \quad (1-26)$$

where A is the cross-sectional area of the junction. Thus the hole diffusion current at each position is proportional to the excess hole concentration at that point. The total hole current injected into the n material at the junction can be obtained simply by evaluating Eq. (1-26) at x_{no}

$$I_p(x_n = 0) = qAD_p/L_p \Delta p_n = (qAD_p/L_p) p_n(e^{qV/kT} - 1) \quad (1-27)$$

The minus sign in Eq. (1-27) means that the electron current is opposite to the direction; that is, the true direction of I_n is in the $+x$ -direction, adding to I_p in the total current. If we neglect recombination in the transition region, we can consider that each injected electron reaching $-x_{po}$ must pass through x_{no} . Thus the total diode current I at x_n can be calculated as the sum of $I_p(x_n = 0)$ and $-I_n(x_p = 0)$. If we take the direction as the reference direction for the total current I , we must use a minus sign with $I_n(x_p)$ to account for the fact that x_p is defined in the $-x$ -

direction

$$I = qA \left(\frac{D_p}{L_p} p_n + \frac{D_n}{L_n} n_p \right) (e^{qV/kT} - 1) = I_0 (e^{qV/kT} - 1) \quad (1-28)$$

Equation (1-27) is the *diode equation* having the same form as the qualitative relation Eq. (1-20). Nothing in the derivation excludes the possibility the total current through the diode for either forward or reverse bias. We can calculate the current for reverse bias by letting $V = -V_r$.

Another simple and instructive way of calculating the total current is to consider the injected current as supplying the carriers for the excess distributions. For example, $I_p(x_n = 0)$ must supply enough holes per second to maintain the steady state exponential distribution Δp_n as the holes recombine. The total positive charge stored in the excess carrier distribution at any instant of time is

$$Q_p = qAL_p \Delta p_n \quad (1-29)$$

The average lifetime of a hole in the n-type material is τ_p . Thus on the average, this entire charge distribution recombines and must be replenished every τ_p sec. The injected hole current (at $x_n = 0$) is needed to maintain the distribution is simply the total charge divided by the average time of replacement

$$I_p(x_n = 0) = Q_p / \tau_p = qA \Delta p_n D_p / L_p \quad (1-30)$$

using $D_p / L_p = L_p / \tau_p$

This is the same result as Eq. (1-26), which was calculated from the diffusion currents. Similarly, we can calculate the negative charge stored in the distribution and divide by τ_n to obtain the injected electron current in the p material. This method, called the charge control approximation, illustrates the important fact that the minority carrier current [for example, $I_p(x_n)$] decreases

exponentially with distance into the neutral region. Thus several diffusion lengths away from the junction, most of the total current is carried by the majority carriers. We shall discuss this point in more detail in the following section.

In summary we can calculate the current at a p-n junction in two ways (a) from the slopes of the excess minority carrier distributions at the two edges of the transition region and (b) from the steady state charge stored in each distribution. We add the hole current injected into the n material $I_p(x_n = 0)$ to the electron current injected into the p material $I_n(x_p = 0)$, after including a minus sign with $I_n(x_p)$ to conform with the conventional definition of positive current in the $+x$ -direction. We are able to add these two currents because of the assumption that no recombination takes place within the transition region. Thus we effectively have the total electron and hole current at one point in the device (x_{n0}). since the total current must be constant throughout the device (despite variations in the current components), I as described by Eq. (1 - 28) is the total current at every position x in the diode.

One implication of Eq. (1 - 28) is that the total current at the junction is dominated by injection of carriers from the more heavily doped side into the side with lesser doping. For example, if the p material is very heavily doped and the n region is lightly doped, the minority carrier concentration on the p side (n_p) is negligible compared with the minority carrier concentration on the n side (p_n). Thus the diode equation can be approximated by injection of holes only, as in Eq. (1 - 27). This means that the charge stored in the minority carrier distributions is due mostly to holes on the n side. This structure is called a $p^+ - n$ junction, where the $+$

superscript simply means heavy doping. Another characteristic of the $p^+ - n$ or $n^+ - p$ structure is that the transition region extends primarily into the lightly doped region. Having one side heavily doped is a useful arrangement for many practical devices, as we shall see in our discussions of switching diodes and transistors. This type of junction is common in devices which are fabricated by counterdoping. For example, an n-type Si sample with $N_d = 10^{14} \text{ cm}^{-3}$ can be used as the substrate for an alloyed or diffused junction. If the doping of the p region is greater than 10^{19} cm^{-3} (typical of alloyed junctions), the structure is definitely $p^+ - n$, with np more than five orders of magnitude smaller than pn. Since this configuration is common in device technology, we shall return to it in much of the following discussion.

In this discussion of carrier injection and minority carrier distributions, we have primarily assumed forward bias. The distributions for reverse bias can be obtained from the same equations, if a negative value of V is introduced. For example, if $V = -V_r$ (p negatively biased with respect to n), we can approximate Eq. (1 - 24) as

$$\Delta p_n = p_n [e^{q(-V_r)/kT} - 1] \approx -p_n \text{ for } V_r \gg kT/q \quad (1 - 31)$$

And similarly $\Delta n_p \approx -n_p$.

Thus for a reverse bias of more than a few tenths of a volt, the minority carrier concentration at each edge of the transition region becomes essentially zero as the excess concentration approaches the negative of the equilibrium concentration. The excess minority carrier concentrations in the neutral regions are still given by Eq. (1 - 25), so that depletion of carriers below the equilibrium values extends approximately a diffusion length beyond each side of the transition region. This reverse-bias depletion of

minority carriers can be thought of as *minority carrier extraction*, analogous to the injection of forward bias. Physically, extration occurs because minority carriers at the edges of the depletion region are swept down the barrier at the junction to the other side and are not replaced by an opposing diffusion of carriers. For example, when holes at x_{no} are swept across the junction to the p side by the ϵ field, a gradient in the hole distribution in the n material exists, and holes in the n region diffuse toward the junction. It is important to remember that although the reverse saturation current occurs at the junction by drift of carriers down the barrier, this current is fed from each side by diffusion toward the junction of minority carriers in the neutral regions. The rate of carrier drift across the junction (reverse saturation current) depends on the rate at which holes arrive at x_{no} (and electrons at x_{po}) by diffusion from the neutral material. These minority carriers are supplied by thermal generation, and we can show that the expression for the reverse saturation current, represents the rate at which carriers are generated thermally within a diffusion length of each side of the transition region.

1.4 Summary

The resistivity of a good conductor is of the order of 10^{-6} ohm $\cdot$ cm, that of a semiconductor, in the range from 10^{-3} ohm $\cdot$ cm to 10^6 ohm $\cdot$ cm, and that of an insulator, from perhaps 10^{10} ohm $\cdot$ cm upward. Bulk properties of a material are properties that are free of perturbations by surfaces.

An oscillator possesses a characteristic frequency, or mode of motion, and a coupled assemblage of N identical such oscillators exhibits N characteristic modes of motion, a phenomenon known

as splitting. An electron in an isolated atom resides in a discrete state that can accommodate only one electron. When two atoms interact by virtue of proximity, each such state splits into two states. In a silicon crystal containing N atoms, each such state splits into N states, where N is typically an enormous number. The splitting phenomenon thus gives rise to so many states “packed into” an energy range of the order of eV in height, that effectively there exists a continuum of allowed states, termed a *band*. Since bands originating from a number of the discrete states that exist in the isolated atom can overlap in the case of a solid, the result is an even denser continuum of allowed states. In semiconductors and insulators, there are bands separated by a gap, the bandgap, in which no (or very few) states exist. Further, the number of states in the lower of these two bands is precisely equal to the total number of outer-shell electrons in all of the atoms of the solid. In an insulator, the gap amounts to several eV (e.g. 5 eV or more), so that the upper band is empty and there are no carriers of electricity in the solid. In a semiconductor, the gap is smaller (e.g. 1 eV), so a few electrons at room temperature can reach the upper band from the lower band. These upper-band electrons and the “holes” they have left behind in the lower band lend modest conductivity to the solid. The upper band is the conduction band; its electrons are conduction electrons, or simply electrons. The lower band is the valence band; its missing valence electrons are *holes*. At low temperature, the electrons revert to the valence band, eliminating the holes and “disappearing” themselves as a result, so that the semiconductor is converted to an insulator. A conductor, by contrast, has a band that is partly filled at any temperature, and is thus capable of conducting at any

temperature.

The probability that an electronic state in a solid is occupied by an electron is a function of the energy position of the state; the relevant expression is known as the *Fermi-Dirac probability function*. This function is symmetric about a point at which probability equals 0.5, a point that defines an energy known as the Fermi level. Above the Fermi level, the occupancy probability (of a state by an electron) approaches zero, below it, unity; for this reason, the Fermi level is a kind of “top surface” of the electron distribution, analogous to the top surface of liquid in a container. A simple-exponential function approximates the Fermi-Dirac function for any energy greater than $4 kT$ above the Fermi level. A different exponential function approximates it for any energy more than $4 kT$ below the Fermi level. These two exponential expressions are known as the Boltzmann approximations to the Fermi-Dirac function. Because of the “strength” of these exponential functions, most of the occupied states in the conduction band (i. e., most of the conduction electrons) are located close to the conduction -band edge, and most of the holes in the valence band are close to the valence-band edge, provided in both cases that the Fermi level lies between the band edges and more than $4 kT$ from either. Because the carriers (electrons and holes) reside close to the band edges, one can define equivalent densities of states N_c and N_v for the conduction and valence bands, respectively, by assuming all available states in the bands to be located right at the band edges. This is known as the equivalent-density-of-states approximation. The densities N_c and N_v are constants at a particular temperature (and exhibit only weak temperature dependence), provided the Fermi level is

restricted to the energy range cited just above. Although N_c is larger than N_v by a factor lying between one and two, for many purposes one can safely assume that $N_c = N_v$. This is known as the *band-symmetry approximation*. With band symmetry assumed, and because of the actual symmetry of the Fermi-Dirac function, the electron density n and hole density p are equal when the Fermi level is at the center of the gap, and the common density value is known as the intrinsic density, n_i . Because of the symmetry of the Fermi-Dirac function, the product of the hole and electron densities at a particular temperature is a constant, independent of Fermi-level position, a statement known as the law of mass action, $pn = n_i^2$. In perfectly pure, or intrinsic, silicon, and with band symmetry assumed, the Fermi level resides at the center of the gap, an energy position often termed the intrinsic level.

Impurities are intentionally introduced into a semiconductor crystal to adjust, its electrical properties, in a process termed "doping". The impurities are "dopants". Such atoms are introduced substitutionally, which is to say, in place of silicon atoms. An atom from Column V of the periodic table is used for doping silicon, with phosphorus being the most commonly used such impurity. Having five outer-shell electrons, these atoms ensure that a neutral silicon sample that includes them will contain more electrons than an intrinsic sample. As a result, the Fermi-level position rises above the center of the gap, just as the top surface of liquid in a container rises when more liquid is added to it. The chosen impurity atoms, fixed in the crystal at randomly positioned sites, introduce localized states within the bandgap. Such impurity atoms and their states can be modeled quantitatively as a hydrogen atom immersed

in a dielectric medium. An electron residing in such a state is likened to an electron "in orbit" about a proton. Its binding energy, however, is smaller than that of the electron in hydrogen by a factor of the dielectric constant squared, and so the phosphorus state is located close to the conduction-band edge; an electron driven out of the state becomes a conduction electron (the analog of a free electron). At room temperature, thermal energy in the crystal typically dislodges electrons from over 99% of these impurity states, so that each atom "left behind" carries a charge $+q$, or is ionized. The customary assumption of 100% ionization means that the Column-V impurity atoms have all "donated" their fifth electrons to the conduction band; hence such impurities are termed donor impurities. Near-100% ionization also causes electron density in a donor-doped sample to equal donor density; these densities typically exceed the density of thermally generated electrons (the only ones present in an intrinsic sample) by a large factor.

Substitutional Column-III impurities, with boron being by far the most common, also introduce localized states in the silicon bandgap. This occurs because of boron's need for one additional electron to complete the kind of bonding (covalent) that exists in a tetrahedral structure. The needed electron will with high probability be stolen or "accepted" from a neighboring covalent bond. An electron so appropriated is bound in place almost as tightly as a normal valence electron in a covalent bond, this because of the great affinity of electron pairs of opposite spin in such bonds. This fact accounts for the positioning of the acceptor state just above the valence band. Such action causes, on average, one hole to be introduced into the valence band for each

boron atom, and also causes over 99% of these atoms to possess an ionic charge of $-q$, attributable to the fourth electron that was captured. In the unlikely event that a hole is "in orbit" about such a negative charge center (which is to say, that the boron state is nonionized), then an "upside-down" hydrogen model can be applied because charge signs are now reversed and because holes "fall up", somewhat like a bubble in a liquid.

When both donor and acceptor states are present and the Fermi level lies between them (the usual case), the acceptor states are likely to be occupied and the donor states, empty, which is to say, that all are likely to be ionized. In a neutral sample (again, the usual case) with more donors than acceptors, one can consider that each acceptor has received an electron from a donor atom, in the process eliminating the latter as an effective donor atom because it now has no electron to donate to the conduction band. In the reverse situation, each donor atom has contributed an electron to an acceptor state, rendering it unable to accept an electron from an adjacent bond and thus to create a mobile hole. This effective cancellation of a minority impurity by a majority impurity is termed compensation. In the case of equal donor and acceptor densities we have compensated-intrinsic silicon.

The most common state of affairs involves a sample of silicon that exhibits overall neutrality. In such a sample the total of positive species, holes and donor atoms (the latter assumed fully ionized), equals the total of negative species, electrons and (fully ionized) acceptor atoms. The resulting expression is a more exact version of the net doping laws that apply to compensated samples, $n \approx N_D - N_A$ and $p \approx N_A - N_D$. Because electron population in a

particular small energy range is fixed by a product of available-state density and occupancy probability in that vicinity, the effects of changing doping and temperature can be illustrated by means of a “computer” that performs an analog multiplication.

A conversion from electron energy ξ to electrostatic potential Ψ can be accomplished (in a band diagram, for example) by choosing consistent references for the two variables and replacing each value of ξ by $\Psi = \xi(-q)$. The direction of increasing Ψ is then opposite to that for ξ because of the electron's negative charge. It is particularly advantageous to choose the Fermi level as reference and the intrinsic (midgap) potential as the potential Ψ , because compact and symmetric expressions for n and p as functions of n_i and Ψ result.

The electron and hole populations in a silicon sample at room temperature are in violent random motion, manifesting their shares of available thermal energy. The average value of the “thermal velocity” associated with this violent motion is a bit over 10^7 cm/s. When a net motion of a carrier population is superimposed on its random motion, the phenomenon is called carrier transport. Such transport can be caused by the existence of a gradient of electrostatic potential, carrier density, temperature, or any combination of these. We remove temperature gradients from consideration by confining attention to isothermal, or nearly isothermal, problems. The existence of an electric field (a potential gradient) causes drift transport. For field values up to about 3 kV/cm in silicon, drift velocity is proportional to field, with the constant of proportionality being known as (hole or electron) mobility. Electron mobility in silicon is roughly three times greater than that for holes. At higher field values, mobility is

no longer constant but becomes a declining function of field. Drift velocity of electrons in silicon levels off ("saturates") at 10^7 cm/s, a value a bit less than average room-temperature thermal velocity, this occurring at 50 kV/cm. Hole mobility shows similar high-field nonconstancy, but the saturation of hole drift velocity is less pronounced than that of electrons.

One factor fixing mobility at a particular field value is temperature, which determines available thermal energy, and hence the density and average energy of vibrational wave packets, or phonons, in the crystal. Two categories of phonons are capable of causing carrier "scattering", or deflection. Acoustic phonons arise from vibrations at sound-wave frequencies in the crystal. Optical phonons arise from vibrations at much higher infrared frequencies. Both kinds of vibrations can be either longitudinal or transverse. Some of the transverse optical phonons in a crystal of zincblende structure also interact strongly with photons.

Other factors affecting mobility are impurity-ion density, with donors and acceptors being equally effective as scattering centers, and screening. The last factor is the tendency of majority carriers to "cluster" slightly near a charge center of opposite sign, reducing the effective range at which its charge can be "felt". In a compensated sample, diminished screening and increased scattering-center density cause especially low mobility. For both carriers, mobility is constant in an uncompensated sample up to a doping density of about $10^{14}/\text{cm}^3$, which is to say that phonon scattering dominates at room temperature in purer samples.

The three requirements for drift transport are (1) a charged species (2) that is in random motion, and (3) that is in the presence of an electric field. For diffusion transport, by contrast,

there are only two requirements: (1) a species in random motion, (2) which exhibits a density gradient. Diffusive flux is proportional to density gradient, a relationship known as Fick's first law. The constant of proportionality is the diffusivity. Of great significance is the fact that drift transport is a function of two variables, carrier density and electric field, while diffusion transport is a function of but one variable, density gradient. This state of affairs is responsible for subtle relationships in device analysis. For a particular population, the ratio of the two transport mechanisms of interest (that is, the ratio of diffusivity to mobility) is equal to the thermal voltage kT/q , an equation known as the Einstein relation.

The density of excess carriers is the difference between actual and equilibrium densities for the species in question. It can be positive or negative. In the most common circumstances and for a neutral sample, excess-electron and excess-hole densities are equal. For this reason, a many-fold change in minority-carrier density can result in a negligible relative change in majority-carrier density. When the majority-density change is significant, the high-level regime has been entered. In the graphical presentation of excess-carrier populations under low-level conditions, a logarithmic scale emphasizes the important relative information that $n_N \approx n_{ON}$, or $p_p \approx p_{op}$, while a linear scale conveys the absolute information that $p' = n'$. The latter choice, however, requires introducing an enormous gap in the scale. Both the band-to-band and recombination-center mechanisms predict low-level recombination rate to be proportional to minority-carrier density. The latter mechanism predicts the same result for high-level conditions, a result confirmed experimentally for the case of

silicon, thus indicating the dominance of this mechanism in silicon. Electronic states that are localized spatially and far from the band edges in energy are termed deep states, and can be produced in silicon by using heavy-metal atoms for doping. The term trap is also used to identify deep states. Such states are effective recombination centers, their effectiveness increasing with nearness of the state to the gap center. Gold, which has an acceptor state 0.54 eV from the conduction-band edge, is particularly effective. Crystalline imperfections that inevitably accompany conventional doping also serve as recombination centers, and the density of such centers increases with total doping, causing carrier lifetime to decline. Typical lifetime values in silicon range from 1 ns to 10 μ s.

The rate of change of the density of a carrier at a point (or the accumulation of carriers in a region) can be equated to the combined effects of transport, generation, and recombination at that point (or near that region), with the result being a continuity equation. Its importance is as a starting point for analysis. A four-term form known as the constant-E continuity-transport equation is particularly useful, but is based upon a significant number of simplifying assumptions. Typical problems require the retention of only two terms. The Haynes-Shockley experiment, however, satisfies all the simplifying assumptions, and yet requires retention of all four terms. It observes the drift of a pulse of minority carriers in a constant electric field, and in principle yields a measure of drift mobility. The kind of mobility cited earlier is conductivity mobility. The two are equal under low-level conditions. A subtle and important facet of low-level conditions is that the drift and diffusion of majority carriers can coexist in a

certain region, while minority carriers in the same region exhibit comparable diffusion, yet (practically) zero drift.

In a steady-state situation wherein minority-carriers diffuse toward a surface to recombine, their diffusion velocity $\sqrt{D_p/\tau}$ for holes) is termed the surface recombination velocity s for that surface. An electrical contact (to a semiconductor crystal) that embodies a high- s surface can meet the two criteria that define an ohmic contact—low resistance and avoidance of carrier-density disturbances. A better kind of ohmic contact, however, is a high-low junction, which is to say a p^+p junction for the case of a P-type region and an N^+N junction for the case of an N-type region.

Steady-state conditions and nonequilibrium conditions can coexist. But equilibrium conditions are a special case of steady state, falling outside the region of commonality. By the opposite token, time-varying conditions are a special case of nonequilibrium, also falling outside the region of commonality.

A PN junction is an interface that separates a P-type region from an N-type region in a semiconductor single crystal. A perfectly abrupt transition from one doping value to another, with uniform doping on each side, constitutes a step junction. The analytical problem posed by a step junction is closely related to that posed by a metal-semiconductor junction and a semiconductor surface. Ionic space charge exists on both sides of the PN-junction surface (metallurgical junction), with equal and opposite charge per unit of junction area on the two sides, thus forming a double layer or dipole layer. In a band diagram of a junction, either band edge can be taken as a potential map of the structure, but it is convenient and “evenhanded” to choose the Fermi level as potential origin, in which case the midgap trace or “intrinsic”

potential becomes the potential. A constant or "flat" Fermi level throughout the sample guarantees equilibrium. Under such conditions, electron drift precisely balances electron diffusion at any position, as do hole drift and hole diffusion.

The depletion approximation for a step junction assumes that ionic space charge is uniform from the metallurgical junction to a boundary at which there is a sharp transition back to neutrality. With this assumed charge profile, integration of Poisson's equation yields a linear (triangular) electric-field profile, and second integration yields a parabolic potential profile. The integral of the field profile, or the area of the triangle, has potential dimensions and represents the contact potential (or built-in potential drop or diffusion potential) of the junction at equilibrium. The height of the field-profile triangle is peak field E_{0M} , and its base is depletion-layer thickness X_0 , so the geometrical relationships in a triangle show how these three important step-junction properties are related.

The contact potential of a junction cannot be measured with a voltmeter because compensating contact potentials exist at other interfaces in the voltmeter circuit. Forward bias means positive voltage on the P-type side and negative on the N-type side, and causes the field profile to shrink to a smaller congruent triangle, thus shrinking all three of the junction properties just cited. Reverse bias does the opposite, adding to contact potential. The depletion approximation improves with reverse bias, and worsens with forward bias. In any step junction, symmetric, asymmetric, or grossly asymmetric (one-sided), depletion-layer thickness X goes as the square root of total potential difference, $\Delta\psi = \Delta\psi_0 + V_{NP}$.

Boltzmann quasiequilibrium assumes that a Boltzmann relation holds for carrier densities at the boundaries of a junction under bias, with $\Delta\psi$ replacing $\Delta\psi_0$. It leads to the law of the junction, which says that under low-level forward bias, a boundary value of minority-carrier density (either side) is given by the equilibrium density multiplied by a Boltzmann factor $e^{-U_{NP}}$, where $U_{NP} \equiv qV_{NP}/kT$. The relevant boundaries in a forward-biased junction are well defined, as surfaces where electric field changes sign. Because minority-carrier boundary values are inflated by forward bias, minority carriers are injected into both neutral end regions. In a one-sided junction, injection into the lightly doped side grossly exceeds injection into the other side. Through a wide range of forward voltage, minority-carrier-density boundary value, density gradient at the boundary, and hence drift current vanish—all rise exponentially with voltage. The resulting I-V characteristic is linear on a semilog plot, but on a linear plot exhibits an offset voltage which in turn decreases as energy gap increases. This simple diffusion picture is complicated for silicon by recombination-generation current. In forward bias, carriers flood into the space-charge region and recombine there, and hence are not injected. At low voltages this current component swamps out the diffusion component (carrier-injection component) of current. In reverse bias, carriers generated in the space-charge layer swamp the “saturating,” or four boundary-value unknowns instead of the two unknowns of minority density, and current increases more slowly with voltage than at lower values.

A model can be a set of equations or an equivalent circuit. In both cases there is a trade-off between accuracy and simplicity, and utility is greatly enhanced by simplicity. Both kinds of models

permit us to understand the relationships of variables and parameters. A parameter is a "temporarily constant variable." Such models can be static (meaning for dc or near-dc conditions) or dynamic (meaning for time-varying conditions). Dynamic models can deal with sinusoidal or switching-transition variations in large-signal in nature. In the small-signal case the I - V or Q - V characteristic can be considered linear without serious error, and such modeling is sometimes described as linearization. In the large-signal case the inevitable nonlinearities of semiconductor devices must be addressed. A piecewise-linear model, which simulates a nonlinear characteristic with a set of straight-line segments, is one way of attacking some features of such a problem. An ideal rectifier in the equivalent-circuit sense has a piecewise-linear characteristic with $R = 0$ in forward bias and $R = \infty$ in reverse bias, and with these lines superimposed upon the current and voltage axes, respectively. In the static charge-control model, the stored charge Q_s of diffusing minority carriers must be replaced "every τ seconds" so that the current of those carriers is Q_s/τ .

A PIN diode involves two-step junctions, can be designed for high breakdown voltage, and exhibits a voltage-independent depletion-layer capacitance, thus simulating an ordinary capacitor. In a linearly graded junction, the depletion approximation predicts a triangular charge-density profile on each side, and improves in accuracy with increasing doping-density profile on each side, and improves in accuracy with increasing doping-density gradient. Junction formation by solid phase diffusion involves delivering impurity atoms to the surface of a semiconductor sample while it is held at an elevated temperature (1 100 °C for silicon). The

diffused-impurity profile depends in detail upon boundary conditions in the process, but normally involves a high density near the surface and a vanishing density well away from the surface. The surface at which this profile intersects a preexisting opposite-impurity profile depends in detail upon boundary conditions in the process, but normally involves a high density near the surface and a vanishing density well away from the surface. The surface at which this profile intersects a preexisting opposite-type net-impurity profile is the location of the diffused PN junction. A high-low junction is a grossly asymmetric step junction with same-type doping on both sides. It has a small contact potential and carries large currents of either polarity with small applied voltages, thus serving as a valuable ohmic contact. Its space charge consists of ions on the high side and surplus majority carriers on the low side in the form of an accumulation layer. In the simplest cases, surplus carriers constitute an example of equilibrium nonneutrality, and excess carriers, a case of nonequilibrium neutrality.

The most important breakdown phenomena are nondestructive and repeatable, and cause a sharp increase in current through a device with a small increase in applied voltage. Avalanche breakdown is the result of impact ionization, or the breaking of a bond by a carrier that has extracted enough energy ($\sim \xi_G$) from a large electric field to eject an electron from a covalent bond, thus elevating it from the valence band to the conduction band. It requires (1) sufficient field, and (2) sufficient "running room", or distance over which that field obtains. It is not simply a critical-field phenomenon. The new carriers so produced also acquire energy from the electric field, and subsequent bond breaking by

these carriers is the cause of the avalanche feature. For all diffused junctions, graded junctions, graded junctions, and step junctions of any variety, depletion-layer thickness at the breakdown voltage is about the same for PN junctions of a given breakdown voltage.

The phenomenon of tunneling causes reverse breakdown at a low voltage (a few volts or less) in a junction with heavy doping on both sides. This occurs because empty conduction-band states on the N-type side reside at about the same energy as filled valence-band states on the P-type side, and the two sets of states are physically very close (separated by a few tens of angstroms), with the barrier being the essentially state-free energy-gap region. Thus the prerequisites for quantum mechanical tunneling are satisfied—a thin barrier and energy-matched states, filled on one side of the barrier and empty on the other. With even higher doping, tunneling can occur in the opposite direction under forward bias through a small voltage range, leading to the incremental negative resistance of the tunnel diode. In such a device, a linearly graded model is more accurate than a step-junction model.

Punchthrough “breakdown” occurs in a diode with a pair of parallel PN junctions separated by a thin layer. The depletion layers interact within the thin layer, causing that of the reverse-biased junction to “pull down” that of the forward biased junction. As a result, current increases exponentially with voltage through several decades of current (at a voltage that is designable), constituting a case of thermionic emission.

Electric field as a function of potential in a step junction at equilibrium can be obtained exactly by analytical methods. The

second integration of the Poisson-Boltzmann equation needed to obtain potential, as a function of position requires numerical methods. The redundancy in either function is enormously diminished by spatial normalization using a scaling length known as the Debye length. At a particular temperature this is a function of sample permittivity and net doping and is a measure of screening distance, or the distance at which a charge can be “felt” or “seen” in the sample. The field and potential spatial profiles both exhibit asymptotes of simple functional form for the deep-depletion region and for the nearly neutral region. Intervening “bridge” expressions also of simple form can be written, and these expressions together with asymptotic expressions collectively constitute a depletion-approximation replacement (DAR), or an approximate-analytic description of these junction properties that is about as accurate as a numerical solution and about as easy to use as the depletion approximation. Analogous treatment can be accorded to other junction properties. A one-sided junction possesses an inversion layer made up of surplus carriers. It is a thin, dense layer of carriers adjacent to the metallurgical junction on the lightly doped side, consisting of carriers that have “spilled over” from the heavily doped side.

When a current or voltage excursion imposed on a diode (or other device) is so small that the relevant portion of its grossly nonlinear I-V characteristic can be approximated by a straight line, then small-signal analysis (sometimes described as linearization of the problem) is applicable. Extremely slow current or voltage excursions are termed quasistatic, and are ideal for defining the small-signal properties of a PN junction. The first of these is the small-signal conductance g of a forward-biased junction, or

dI/dV_{NP} . It is given through several decades of current for any semiconductor material. Also, this relation holds for both long and short diodes, but current at a given voltage is larger in the latter. Placing a resistor $r = 1/g$ in series with a voltage source yields a useful piecewise-linear model for the forward-biased junction.

In low-frequency diffusion capacitance, positive and negative charges are stored in the same physical volume in the form of excess carriers. This charge, and hence diffusion capacitance dQ_s/dV_{NP} increases exponentially with forward voltage. Depletion-layer capacitance dQ_s/dV_{NP} involves the storage and recovery of majority carriers at the boundaries of the space-charge layer. Analysis based upon the depletion approximation yields good approximate expressions for depletion layer capacitance. It typically exhibits a power-law dependence upon voltage, dominating in reverse bias and low forward bias, but being swamped by diffusion capacitance at larger forward bias. The two capacitances are equal, or "cross" can write a universal relation between depletion-layer capacitance and voltage that is valid for any step junction. The same cannot be done for diffusion capacitance.

Diffusion, drift, and displacement currents all can play a role in dynamic junction phenomena. A useful equivalent-circuit model of the junction diode consists of the conductance g , the diffusion capacitance C_s , and the depletion-layer capacitance C_t in a parallel combination.

In charge-control analysis, a dynamic extension of static charge-control analysis is assumed. The basic charge-control assumption concerning diffusion capacitance is that the spatial distribution of charge under dynamic conditions is the same as

under static conditions. Since this is significantly inaccurate, one often employs an effective capacitance, accompanied by an effective carrier lifetime. Useful approximate results are often obtained by letting the effective values be half those defined quasistatically.

Much charge-control analysis is applied in the upper portion of the low-level forward-bias regime, wherein depletion-layer capacitance can be neglected. Thus one exploits the fact that the dynamic charge-control expression is a linear differential equation, and the principle of superposition can be applied, adding together the solutions of several simple linear problems in order to treat a more complicated problem. For still more general problems, the grossly nonlinear depletion-layer capacitance is sometimes added subsequently to the analysis.

There are at least four techniques or models for analyzing the junction diode in terms of its internal physics. The most accurate, termed exact here for brevity, starts with the time-dependent diffusion equation; this expression is the result of dropping the field-dependent term from the constant-E continuity-transport equation. A second method, finite-difference, divides up the region of minority-carrier storage into many smaller segments, or "lumps," and treats the problem numerically. In the limit of many segments, its accuracy approaches that of the exact model. In the third, usually called the Linvill model, only two or three lumps are used, so that analytical treatment of the problem is possible. The last and least accurate is the charge-control method, where a single lump is employed. Here we have focused on the two extreme methods-charge-control and exact.

Also, we use linear-network methods to treat a sinusoidal

input signal and a step or switching input. In switching, a diode or transistor is taken quickly from one operating state to another. This binary application of such devices has enormous practical importance. It has been the basis of electronic computing since its beginning in the 1930s and 1940s, and is now encroaching steadily into applications that previously were exclusively "analog," or amplifierlike.

A device-physics analysis of the diode that starts with the diffusion equation and combines this with the charge-control model permits derivation of the linear differential equation for current as a function of time-dependent stored charge, as well as voltage and conductance as functions of charge. Solving the current equation for the case of a small-signal current step yields a circuit-behavior solution with an exponential voltage waveform involving effective lifetime as the characteristic time. Circuit-theory analysis confirms the result. Converting to a sinusoidal source permits determination of diode admittance. Generalizing to the large-signal switching problem yields a differential equation of the same form as that for the small-signal problem, and hence a solution of the same form as well. Generalizing further to the case of current turn-off, or the reversal from forward to reverse current, one can obtain an estimate of the storage time required to remove stored minority carriers, a time that is dependent on the ratio of the reverse-current to the forward current magnitude. The model predicts unrealistically an abrupt transition to zero current at the end of the storage time (The more realistic exact analysis predicts a decay time following storage.) In the opposite case, turn-on, effective lifetime again determines the exponential response.

Device-physics exact analysis starts with the time-dependent

diffusion equation and determines the time-dependent spatial distribution of charge during large-signal switching. From this it infers storage time as the time required at constant current to reduce stored-charge density at the junction boundary to zero. Thereafter it uses an approximate method to determine current versus time during removal of the remaining stored charge, and infers (current) decay time from this function. Analysis of diode response to a small-signal sinusoidal applied voltage also starts with the time-dependent diffusion equation. The resulting carrier profile can be described as an exponential static component with an extremely heavily damped sinusoidal dynamic component superimposed upon it. Perturbation theory permits the writing of static and time-dependent current components, as well as diode conductance and admittance. The general admittance expression can be readily approximated for high and low frequencies, and for diodes of various lengths.

Using circuit-behavior exact analysis, we present four solutions corresponding to small-signal sinusoidal and positive-going step stimuli, and current and voltage sources. For comparison these are augmented by four corresponding solutions based upon charge-control analysis. It is shown that the small-signal current-source step and sinusoidal responses are linked by the fact that the product of rise time and bandwidth is a constant—with a value of 0.29 from the exact model, and 0.35 from the charge-control model. As a result, time-domain response can be determined from the frequency-domain response, and vice versa. Eight corresponding and significantly more complicated large-signal solutions are also given, for a total of sixteen cases.

The SPICE program for junction-diode switching simulation

focuses on terminal properties and is based upon an empirically modified charge-control analysis. It is capable of incorporating voltage-dependent transition-region capacitance in addition to diffusion capacitance, thus producing a nonlinear dynamic charge-control equation. Further refinements of the model include parasitic series resistance, the diode quality factor for better description of doping profiles. Understanding of the SPICE model is aided by dividing current into two components: one charging or discharging the capacitances, the other being “resistive” current. The advantage of so doing is that the latter involves identically the current-voltage-charge relationship examined quasistatically, while the former can be handled numerically through repeated application of the $i = CdV/dt$ relationship. The relevant capacitance now is of course the sum of the diffusion and depletion-layer capacitances, both being voltage-dependent. A numerical example tracks these two current components in time, along with the associated charges (as well as the resistive current component), with a result that is surpassingly complex in view of the simple total-current waveform. The normal SPICE program does not deliver information on these “internal” variables (current and charge components); it is necessary to insert additional code into the SPICE program to obtain such data.

NOTES

1. As a consequence, the uppermost electrons in the partially filled band or electrons at the top of the valence band can move to the next-higher available energy level when they gain kinetic energy (e.g., from an applied electric field).

结果,半满带中的最上层电子以及价带顶部的电子在获得动

能(如施加的外电场)时可以运动到与其相应的较高能级上。

2. In a band diagram of a junction, either band edge can be taken as a potential map of the structure, but it is convenient and “evenhanded” to choose the Fermi level as potential origin, in which case the midgap trace or “intrinsic” potential becomes the potential.

在一个 PN 结的能带图中,能带的边缘可以视为此结构的电势图,如选择费米能级作为势能的始点更为方便准确。在这种情况下,带间或本征能级就可视为势能。

New Words and Expressions

semiconductor *n.* 半导体

conductivity *n.* 电导率

resistivity *n.* 电阻率

bipolar transistor *n.* 双极晶体管

rectifier *n.* 整流器

photodiode *n.* 光敏二极管

device *n.* 器件

leakage current *n.* 漏电流

silicon dioxide *n.* 二氧化硅

lattice *n.* 晶格

unit cell 晶胞

facet *n.* 晶面

lattice constant 晶格常数

diamond lattice 金刚石晶格

zincblende 闪锌矿结构

bcc 体心立方

fcc 面心立方

energy levels 能级

Miller indices 密勒指数
 hole *n.* 空穴
 permittivity *n.* 介电常数
 covalent bonding 共价键
 conduction band 导带
 effective density of states 有效态密度
 intrinsic *adj.* 本征的
 extrinsic *adj.* 非本征的
 carrier *n.* 载流子
 bandgap *n.* 能带间隙
 mass action law 质量作用定律
 donor *n.* 施主
 acceptor *n.* 受主
 injection *n.* 注入
 impact ionization 碰撞电离
 scatter *v.* 散射
 drift *v.* 漂移
 mean free path 平均自由程
 mean free time 平均自由时间
 mobility *n.* 迁移率
 diffusivity piecewise 分段扩散率
 recombination *n.* 复合
 diffusivity *n.* 扩散系数
 decay *v.* 衰减
 abrupt *adj.* 突变的
 transition *n.* 跃迁
 step junction 突变结
 metallurgical junction 合金结
 Fermi-level 费米能级

exponential *adj.* 指数的; 幂数的
dopant *n.* 掺杂
dielectric *n.* 电介质; 绝缘体
dislodge *v.* 移出; 移走
donor *n.* 施主
bonding *n.* 键合
tetrahedral *adj.* 四面体的; 有四面的
phonon *n.* 声子
packet *n.* (信息、数据)包
acoustic *adj.* 声学的
lifetime *n.* 寿命
continuity equation 连续性方程

参 考 文 献

- 1 S M Sze. Semiconductor Devices. Bell Telephone Inc., 1985
- 2 R M Warner, B L Grung. Semiconductor-Device Electronics. 北京: 电子工业出版社, 2002
- 3 Ben G Streetman. Solid State Electronic Devices. Prentice-Hall, Inc., 1980

Chapter 2

Semiconductor Device

2.1 Bipolar Junction Transistor

The bipolar junction transistor (BJT) was invented in January 1948, a few weeks after the invention of the first transistor (the point-contact transistor), and months before the point-contact device was announced. The reason for the invention of the BJT is a prime object lesson in engineering practice. The point-contact transistor was fabricated by primitive and “artistic” methods, and its detailed structure and operation remain unclear to this very day. Even worse, it posed a complex three-dimensional problem. The inventor of the BJT, William Shockley, was at the time the supervisor of the two inventors (or “discoverers”) of the point-contact transistor, John Bardeen and Walter Brattain. He was endeavoring, with sound engineering instinct, to conceive of a one-dimensional analog of the point-contact device as a way of opening the door to analytical investigation of its operation. Suddenly he realized that the structure he was postulating was itself a transistor, and onto that lent itself to analysis at a basic level! It became the bipolar junction transistor. Reduced to practice in 1951, it had for all practical purposes shouldered the point-contact device aside by the mid-1950s. A genuine

understanding of the point-contact transistor will probably never be achieved, because it is inferior to competing devices in so many respects that it no longer arouses even academic interest. Nonetheless, its place in history is assured as the first practical amplifying device in the era of solid-state electronics.

Early in the 1950s, Shockley proposed the adjective bipolar to describe his newly conceived device because both carrier types play important roles in its operation; by contrast, he described devices in the field-effect family (which he also launched insofar as practical embodiments were concerned) as unipolar, because their operation depends primarily on a single carrier type. In more recent times, the need to distinguish among the proliferating varieties of transistors has caused the term bipolar junction transistor to come into common use, and the resulting acronym BJT to come into even more common use. This kind of distinction is important. Shockley's first field-effect device was a "unipolar junction transistor," but it is usually designated the junction field-effect transistor. The term "unipolar" is not widely used.

The unfolding story of solid-state electronics can be told rather completely in terms of evolving fabrication technology, constantly expanding the number of options available to the device and integrated-circuit designer. It was for technological reasons that an early and important kind of BJT was a germanium PNP device. The term PNP labels the conductivity types of the three regions within a BJT, regions separated by two PN junctions. In later years, and again partly for technological reasons, the dominant BJT was a silicon NPN device. In integrated circuits today, the combination of silicon NPN and PNP devices is a growing practice because the resulting complementary circuits

have important power-dissipation and performance advantages. For convenience and consistency, however, and because of its continuing importance, the silicon NPN BJT will be the vehicle for this chapter.

2.1.1 Structure and Technology

The essential structure of a BJT is represented in Figure 2 – 1 (a). The very earliest such devices had structures literally of this kind. Two closely spaced junctions were created by crystal-growth methods, and a “bar” or parallelepiped was then cut out of the germanium crystal. Electrical leads were attached to it (an enormous challenge!) and the result was a BJT. For reasons that will be explained shortly, these electrical terminals are given the names, respectively from left to right, emitter, base, and collector. These names were chosen with an eye to distinctive initial letters, which are displayed in Figure 2 – 1(a) in association with the three terminals. The shaded regions in Figure 2 – 1(a) represent the space-charge regions (or depletion regions) of a pair of NP junctions. The boundaries of these regions are emphasized because conditions there assume far greater importance than conditions at the metallurgical junctions (which are not even represented in the drawing).

The term “base” may well be a puzzling choice in the context of Figure 2 – 1(a). Actually, the term is a vestigial remnant (and practically the only remnant) of the point-contact transistor. In that device a small germanium crystal was mounted on a pedestal, and two point-contact wires (designated as emitter and collector) were positioned on top of the crystal, so that the crystal was literally the “base” of the structure. The central region in

Figure 2 – 1(a) plays an electrical role analogous to that of the base region in a point-contact device.

Having given names and single-letter labels to the three terminals of the BJT we have effectively provided names and labels to the three regions of the BJT as well, because each region is associated uniquely with one terminal. The matter of junction identification is almost as simple. The two junctions are, respectively, the emitter-base junction and the collector-base junction. However, without any ambiguity we can (and usually do) refer to them simply as the emitter junction and the collector junction.

The corresponding BJT symbol, Figure 2 – 1(b), was adopted very early. Once again an arrowhead is used to point from a P-type region to an N-type region, the direction of the “easy” conduction of conventional current through a PN junction. But only one junction-the emitter junction-is so labeled. (This is done for identification.) Once the emitter junction has been so identified and labeled, the polarity of the collector junction becomes evident.

In 1960 the BJT took on a structure that it has retained to the present day in a wide range of applications, the structure shown in Figure 2 – 2. The orientation of this diagram has been chosen to correspond to that of the BJT in Figure 2 – 1(a). It is evident that this structure bears little resemblance to the prototypical “bar” structure. Nonetheless, in a large number of devices having the newer structure, the currents through the junctions retain an essentially one-dimensional character, with important current fractions and current components being normal to the junctions. Consequently, a good approximate treatment of such devices can

be achieved by calculating currents in the region emphasized by dotted lines in Figure 2 - 2 and then multiplying by an appropriate junction area to determine total current through the junction. Thus Shockley's original one-dimensional intent is still being honored.

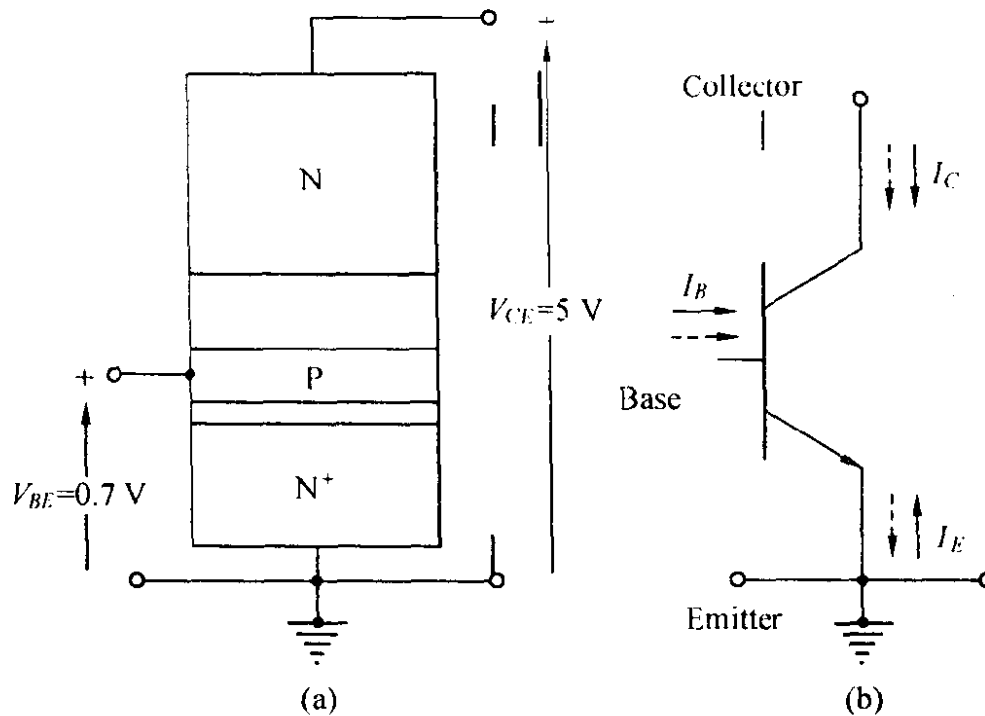


Figure 2 - 1 A representation of the NPN bipolar junction transistor(BJT)

Figure 2 - 2 specifically represents the structure of a discrete or separately packaged BJT. It is evident from the dimensions shown that the active portion of the device is confined to a small portion of the overall crystal volume. The bulk of the silicon crystal, in fact, is simply a mechanical support for the active portion. Any electrical resistance to the passage of current through this thick region is a parasitic feature of the device that one would prefer to avoid. Most BJTs today are incorporated into an integrated circuit (the "microchip" of the layman), defined as a useful combination of devices fabricated in and in some cases on a semiconductor single crystal. As a result, the parasitic feature of having current pass through a thick mechanical support is partly

eliminated. All of the currents (in virtually all integrated circuits) are confined to a very thin region near one face of a semiconductor crystal.

The portions of the BJT structure shown in Figure 2 - 2 that are near the top and the bottom of the diagram also are to some degree "parasitic." (For a more usual orientation of the diagram with the silicon surface horizontal, these regions would be described as lateral portions of the structure.) It is a curious fact that advancing technology is now generating structures that eliminate the parasitic lateral regions, thus increasingly resembling the original bar structure! These BJTs are not freestanding parallelepipeds, but rather are defined by parallel grooves cut into a silicon crystal, and then subsequently filled by an insulating material such as silicon dioxide (SiO_2). The individual BJTs are interconnected to perform useful circuit functions by means of a variety of methods; the result is an integrated circuit that may incorporate over 100 000 devices all fabricated simultaneously in the same silicon crystal.

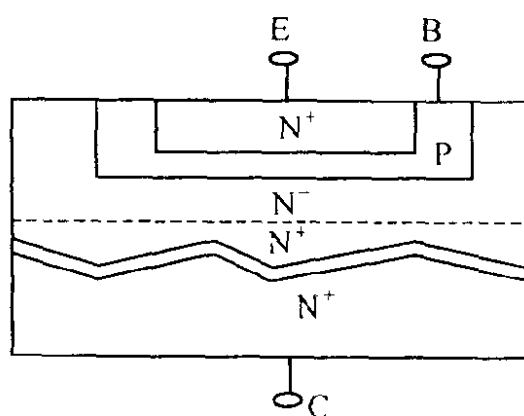


Figure 2 - 2 A BJT structure, employing the epitaxial growth of silicon

2.1.2 Biases and Terminal Currents

An input port or an output port in the circuit sense consists of a pair of terminals. Since the BJT is a three-terminal device, one of the three terminals is permitted to be common to the input and output ports, and the other two terminals are each uniquely associated with each port. For reasons that we shall examine, when the base and emitter terminals are chosen for the input port with the collector and emitter terminals taken for the output port, the BJT can exhibit both current gain and voltage gain. This useful combination of properties has made the common-emitter configuration the most widely used of the several possibilities, a term acknowledging that the emitter terminal is common to the input and output ports. The adjective “grounded-emitter” is also sometimes applied to this connection.

Figure 2 – 1(a) shows the physical representation of the BJT in its common-emitter orientation, and also shows typical voltage polarities and magnitudes for the case of a silicon device. A positive voltage is applied at the input port (with reference to the common-emitter ground potential shown here), and it is evident that the emitter junction is forward-biased as a result. The junction voltage is near 0.7 V. Or, continuing to use the double-subscript notation introduced in Chapter 1, we have $V_{BE} \approx 0.7 \text{ V}$. Collector-emitter bias, or V_{CE} , has a typical value of 5 V in wide ranges of circuits, but for some purposes may be set as low as one volt, or as high as many hundreds of volts.

The BJT symbol in Figure 2 – 1(b) has associated with each terminal a pair of arrows representing current directions. The solid arrows show the directions of conventionally positive terminal

current-always in ward. The dashed arrows show the actual directions of positive conventional current at each terminal for the bias conditions shown in Figure 2 – 1(a). Note that in the cases of base and collector currents, the two most important to the common-emitter problem, the two current conventions are in agreement with respect to direction, a matter of significant convenience.

Operation of the BJT can be approached by examining the properties of its two junctions individually, considering them to be isolated. As we saw in Chapter 1, this is term that describes a junction whose end regions are extensive in the x direction (since we continue to consider one-dimensional structures only). As a result, the carrier-density disturbances that accompany junction biasing have space enough to “fade away” before they reach the inevitable contacts. Thus the inherent properties of the junction under study will be seen. And since we are going to assume step junctions only, these inherent junction properties subsume those of both of its uniformly doped regions-properties such as absolute doping and carrier lifetime. Then we shall combine the junctions in order to examine their interaction. In a BJT, this interaction is crucial; there is no way to simulate BJT behavior using isolated junctions.

In Figure 2 – 3 (a) is shown the BJT structure of interest. The dimensions of the various regions are shown with qualitative realism. The emitter-junction space-charge layer is typically much thinner than that of the collector junction, for two reasons: First, the mean doping of the emitter and base regions (a geometric mean is appropriate here because their net-doping values have a large ratio) is typically almost two orders of magnitude larger than

the mean doping of the base and collector regions; and second, the emitter junction is usually forward-biased, while the collector junction is usually reverse-biased. The thickness of the latter's space-charge region we shall designate as X . The thickness of the base region X_B is typically a few tenths of a micrometer in a modern transistor, and is a critical dimension.

Since the late 1950s, various techniques involving the high-temperature solid-phase diffusion of impurities have been extensively employed in BJT fabrication. In particular it has been emitter regions and base regions that have been formed by these methods, yielding BJT cross sections well represented by Figure 2 – 3. The impurity profiles that result from impurity-diffusion procedures are not the constant-density profiles incorporated in an ideal-step-junction description. Instead they approximate the Gaussian function and the error-function complement as described in chapter 1. As a result, the net-doping-density profiles for both the emitter and base regions exhibit maxima that are skewed toward the surface of the device.

Fortunately for the student and engineer, it turns out that these rather major departures from the ideal structures defined in Figure 2 – 2 are less consequential than one might normally assume. For a combination of reasons, calculations based upon idealized structures can provide meaningful descriptions of device properties and worthwhile guides for design modifications. Therefore, in what follows, we shall continue to adhere to the assumptions of step junctions and uniformly doped regions. It is a curious fact, once more, that modern technology is moving ever closer to the step-junction ideal employed in the simplest analyses, with uniform doping in the emitter and collector regions, and

sometimes in the base region as well.

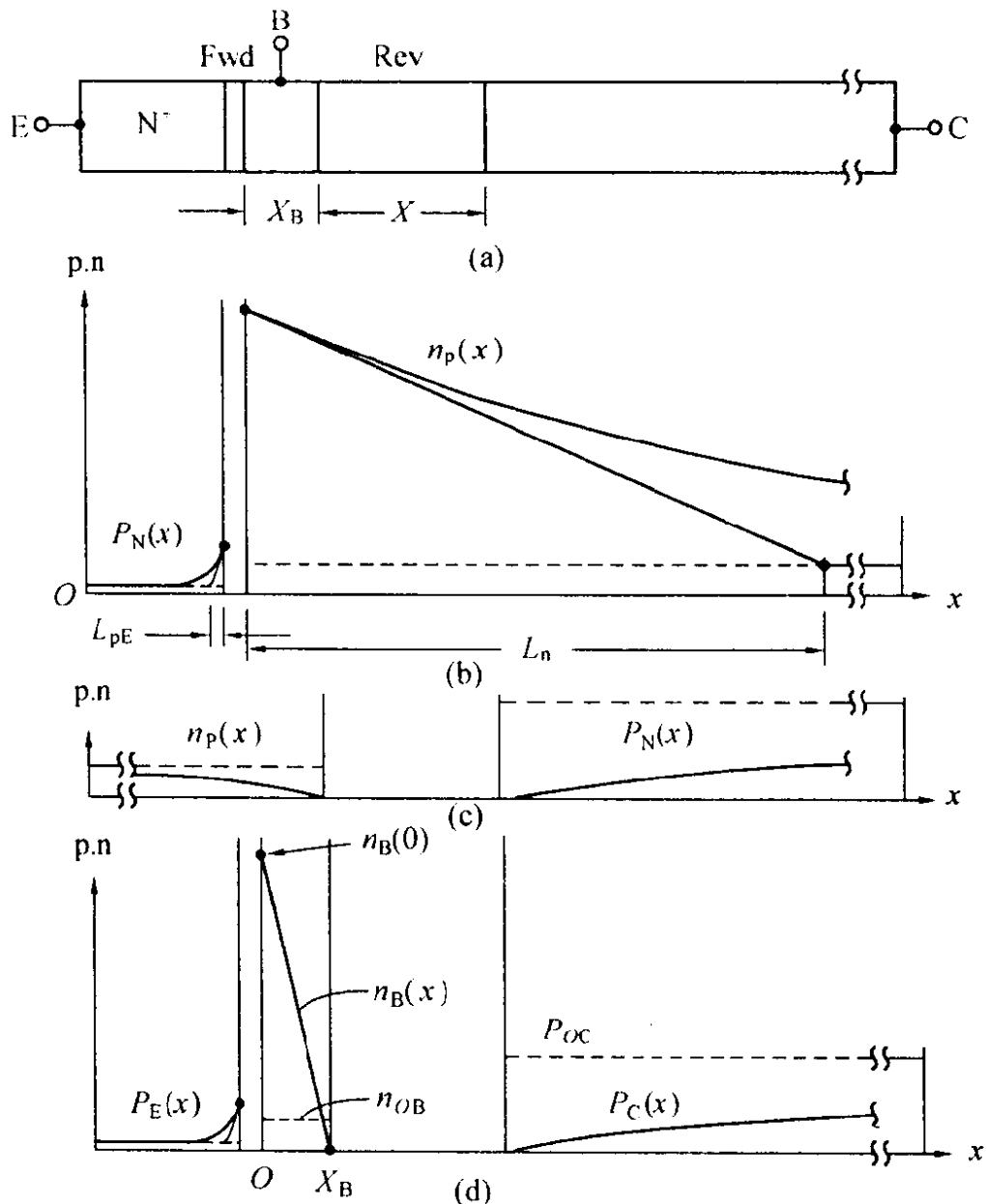


Figure 2 - 3 Minority-carrier profiles in a normally biased BJT

2.1.3 Internal Current Patterns

A detailed examination of current components in the biased BJT reveals a relatively complicated picture. Fortunately, there exist approximations that are simple and yet quite accurate through wide ranges of conditions. This is another way of saying that some

of the most complicated features of the device are often small enough to be neglected safely.

Figure 2 - 4 is a physical representation of a BJT in the "common-emitter orientation" shown first in Figure 2 - 1(a). The large primary arrow directed from collector to emitter is identically, representing one-dimensional electron transport. The labels I_C and I_E show that it constitutes both the collector-terminal current and the emitter-terminal current, which typically differ by only about one percent. Shown also are other current "strands" that depart from the primary current or join it. The relative magnitudes of the secondary currents are represented qualitatively by their widths, but for clarity all of these widths are grossly exaggerated with respect to that of the primary arrow. The circled values give typical currents one might encounter in a small BJT (a "milliwatt device") under bias, and it is evident that they differ by many orders of magnitude.

The most important of the secondary currents is the one labeled I_B , the base-terminal current. It differs in numerous and major ways from the primary electron current that has occupied us up to this point: First, it is a majority-carrier current, or a hole current in the NPN device. Second, it flows laterally from the base contact into the active region of the BJT. Refer back to Figure 2 - 2, which shows the commonplace (diffused) BJT in a relatively realistic cross section. The portion of the base region contiguous with (or "under") the emitter region is the active base region, sometimes also termed the intrinsic base region. (This use of the term does not have doping connotations.) The outer portions of the base region, in analogous fashion, are sometimes designated by the adjective extrinsic. It is there that the base contact is made

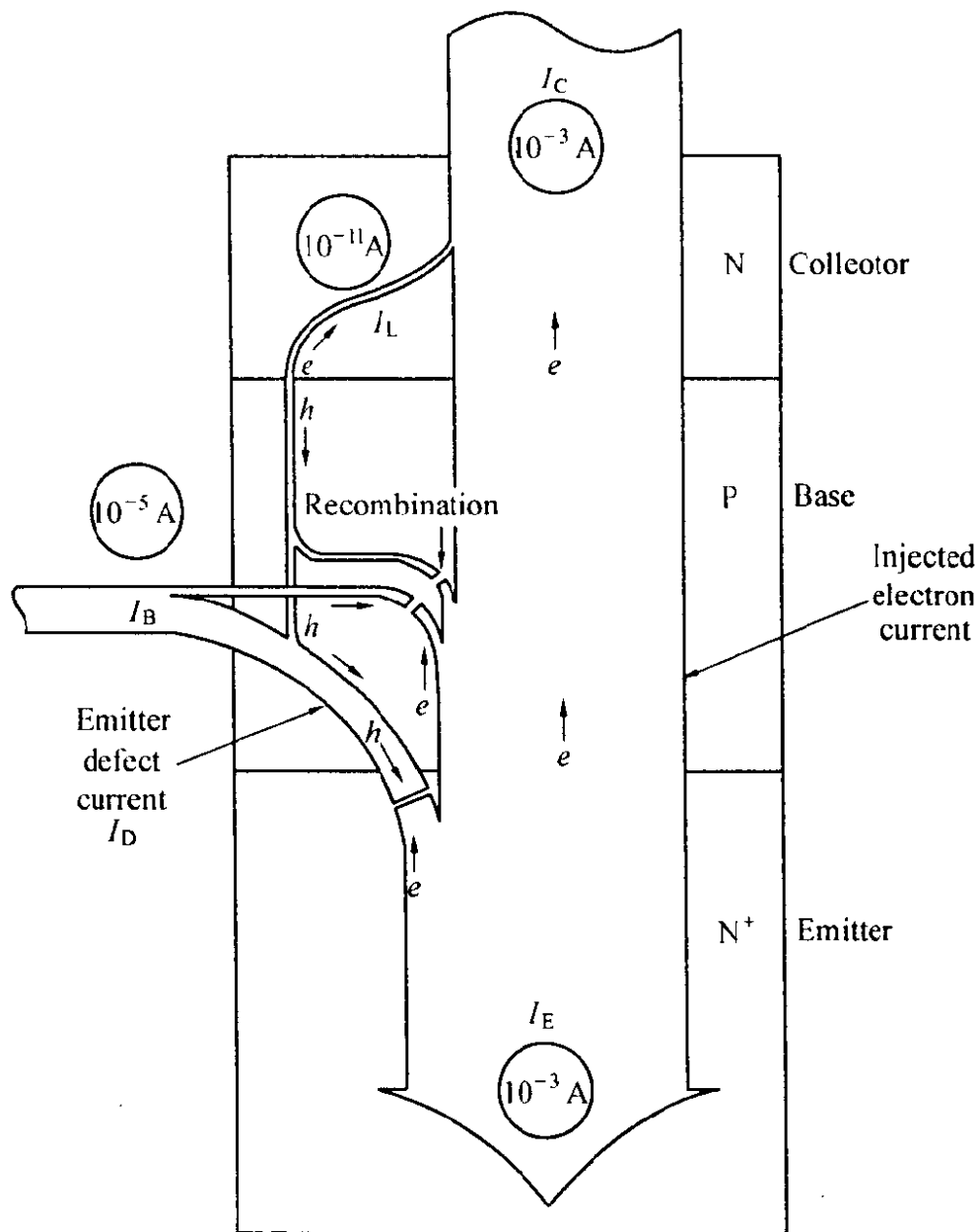


Figure 2 – 4 Representation of electron and hole (e and h) current component in BJT under bias

in this kind of BJT. (Base contacts, by the way, can be and often are made on both sides of the emitter) The third major distinction of the base current from the primary electron current thus becomes evident. The current of holes traverses a relatively long path within the base region. In the example we are considering it could be several tens of micrometers, as compared to the less-than-one-

micrometer base thickness traversed by the electrons.

Once in the intrinsic base region, most of the holes undergo a change in direction and flow into the transition region of the emitter junction. This obviously involves a complex current pattern. It is not a "one-piece" current like that shown for simplicity and convenience in Figure 2 - 4, but rather is somewhat like the pattern shown heuristically in Figure 2 - 5. In spite of the relatively small size of the base current, it is of extreme importance because, as we shall see, it is the control current that is supplied to the input port in the common-emitter BJT. The perturbing effects of the complicated base-current pattern on, for example, the electrostatic-potential pattern in the base region, are not serious so long as the base current is small. In the typical device we are considering, base current amounts roughly to one percent of the electron current, which does qualify as "small".

The resistance encountered by the base current on its lateral path from the (ideal, let us say) base contact to the active base region, and in the active base region as well, plays a key part in fixing device performance. This ohmic base resistance is a parasitic element to be minimized. Its value is typically in the range of 50 to 100 ohms. It is in order to diminish base resistance that the emitter is often given the (plan-view or top-view) shape of an elongated rectangle or stripe, with the base contact, also stripelike, in close proximity to the emitter stripe. Such a structure is shown in Figure 2 - 6. A cross section of this device taken at A-A will approximate the structure shown in Figure 2 - 2. As noted earlier, the BJT designer sometimes chooses to place a base contact on both sides of the emitter stripe; doing so makes a significant reduction in ohmic base resistance.

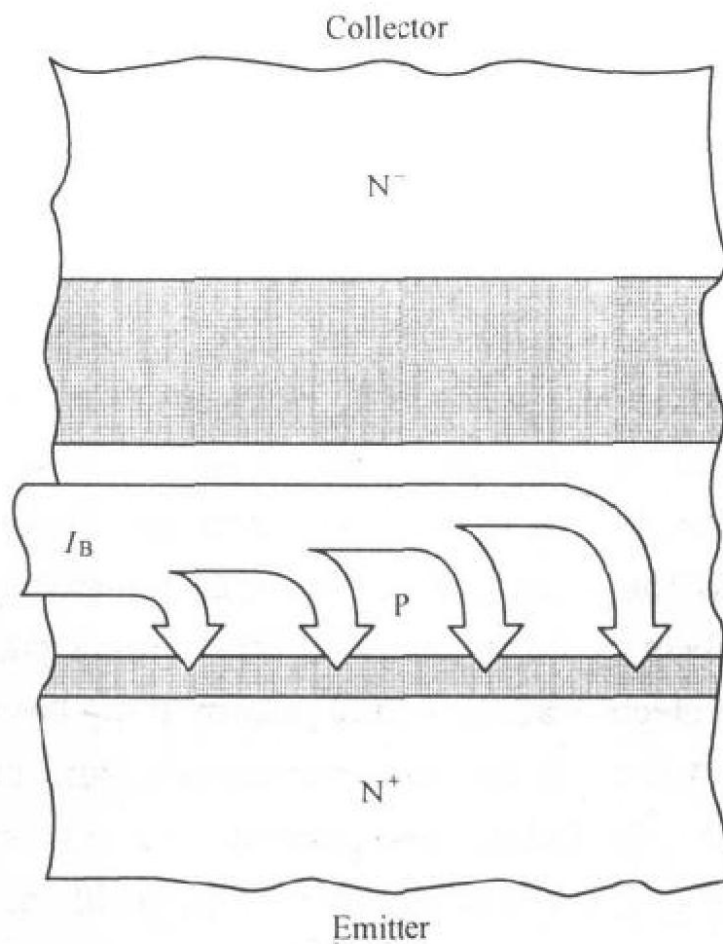


Figure 2 - 5 Heuristic representation of base-current pattern

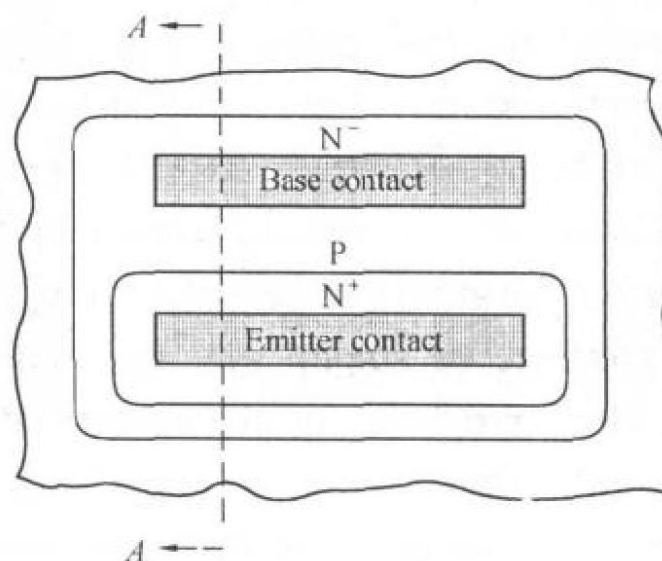


Figure 2 - 6 Typical plan-view geometry for a milliwatt BJT

It is the dimensions of a base or emitter stripe that are

appropriately designated length and width. For dimensions in the x direction, into the silicon, the term thickness is far more descriptive and appropriate, because the latter dimensions are typically much smaller. More important, clarity is served by consistency in these matters. Curiously, even engineers, who are supposed to have above-average skill in spatial visualization, sometimes seem to have difficulty in understanding and verbalizing the three-dimensional aspects of the BJT. One too frequently hears and reads tangled syntax such as. "This transistor has a very thin base width," instead of the simple declaration that the transistor has a thin base region.

Most of the holes shown entering the emitter junction in Figure 2 – 5 pass through it and recombine in the low-lifetime N^+ emitter region, so long as the BJT is not being operated with extremely low current levels. These holes are identically those represented by the profile labeled $P_E(x)$ that is shown at the left in Figure 2 – 3 (d). In Figure 2 – 6, the recombination is symbolized by the gap in the I_B arrow inside the emitter region, with holes approaching it from the base side and the recombining electrons being supplied within the emitter.

2.1.4 Parasitic Internal Currents

The remaining two currents are parasitic in the sense that they vanish in the ideal BJT, and nearly vanish in the silicon BJT. First, a small current is shown in Figure 2 – 4 as splitting from I_B to a recombination gap within the base region. A small but finite portion of the flood of electrons diffusing from emitter to collector fails to reach the collector, recombining in transit, and it is this occurrence that the second gap symbolizes. In the early BJTs,

most of the base current was consumed by this mechanism because the base thickness X_B amounted to tens of micrometers and base-region lifetime values T_B were sometimes low. As a result, the crucial electron profile in the base region departed significantly from the modern-device linear profile shown in Figure 2 – 3 (d). The base-transport efficiency γ_β took on considerable importance in those days. (Today it is so close to unity that it is rarely mentioned.) The factor γ_β can be defined as the ratio of electron current departing from the active base region at its right-hand surface to the electron current entering the active base region at its left-hand surface.

Returning now to the modern transistor, we stress that I_B is very nearly equal to the current of holes being injected into the emitter. This brings us to a second efficiency factor, (γ_E , one that is of great importance in the modern BJT. It can be defined as

$$\gamma_E \equiv (I_{nE}/I_E) = (I_e - I_{pE})/I_E = 1 - (I_{pE}/I_E) \quad (2-1)$$

where I_{nE} and I_{pE} are, respectively, the electron current being injected into the base region by the emitter junction and the hole current being injected into the emitter region by the same junction. The total emitter-junction current is inevitably equal to the emitter-terminal current I_E . In a “perfect” emitter, all of the emitter current would consist of electrons injected into the base region. From this point of view I_{pE} (the current of holes injected into the emitter region) is a defect current, a description sometimes used. The adjective “perfect” was placed in quotation marks, however, because the injected current of holes plays the critical role of the control current, and as such is a vital factor in BJT design and operation, rather than a factor to be eliminated as the potentially misleading term “defect current” may suggest.

This brings us to the last and smallest current, namely the leakage current I_L arising at the reverse-biased collector junction. It provides one of the sharpest contrasts between the modern silicon transistor and the early germanium transistor. As we noted in Chapter 1, the behavior of a reverse-biased silicon junction fits simple theory quite poorly, while the germanium case fits it very well. At the same time, the silicon-junction features that make it nonideal from a theoretical point of view serve to create near-ideal BJT properties! These details are worth examining because of the insight they provide into certain BJT subtleties. A lesser reason is that part of the older literature on the BJT and its applications dwells on the nonideal BJT properties that arise from the “ideal” germanium-junction behavior, and this literature is puzzling to a newcomer without a measure of appreciation of the responsible phenomena.

We saw in Chapter 1 that the dominant reverse-current mechanism in a silicon junction is the excess of generation over recombination in the space-charge region, where reverse bias has caused a depression of both carrier densities below their equilibrium values. (In such conditions, generation exceeds recombination, just as the reverse is true in the presence of excess carriers.) The carriers so generated are separated by the electric field and are swept to the regions where they become majority carriers. A representative value for this leakage current is of the order of ten picoamperes. The description just given shows that it consists equally of holes and electrons. The electrons become a negligible part of the collector current, departing the BJT through the collector terminal. The holes entering the base region mingle with those coming from the base contact. As a result, they, too,

participate in emitter-region recombination and, to a much lesser degree, base-region recombination.

Conditions in the old germanium BJT were quite different. First, the diffusion of minority carriers to the reverse-biased collector junction provided most of the collector leakage current. Because the collector region is more lightly doped than the base region by one to two orders of magnitude, minority holes in the collector region dominate this diffusion mechanism. This density is roughly four million times larger in germanium than in silicon for given net-doping values, because the values of n_i^2 stand in about that ratio for the two materials, and their mobility ratios introduce another factor of about three. The resulting current of holes into the base region might be of the order of 0.1 microampere, amounting to one percent of the $10\mu\text{A}$ base current cited in the example of Figure 2 - 4. When I_B is further diminished (and in particular, when it is reduced to zero), this leakage current is like an irreducible "parasitic base current." Further, this parasitic current is then multiplied by the gain mechanism of the BJT, typically by 100, with very undesirable effects. The details of this topic are best treated after we have introduced and described the gain mechanism itself, which we will do shortly. And after that, we shall point out why the silicon BJT does not multiply its collector-leakage current (already three orders of magnitude smaller) by its current-gain factor.

2.1.5 Common-Emitter Current Gain

In the silicon BJT we may safely neglect for a wide range of purposes the two parasitic internal currents, namely, the base-recombination current and the collector-leakage current. Thus the

admittedly complex pattern of Figure 2 – 4 reduces to the much simpler pattern of Figure 2 – 7. Even here, however, there remains the complexity of the base current's two-dimensional pattern, a complication that may be neglected for a small ratio of base current to electron current, as was noted earlier.

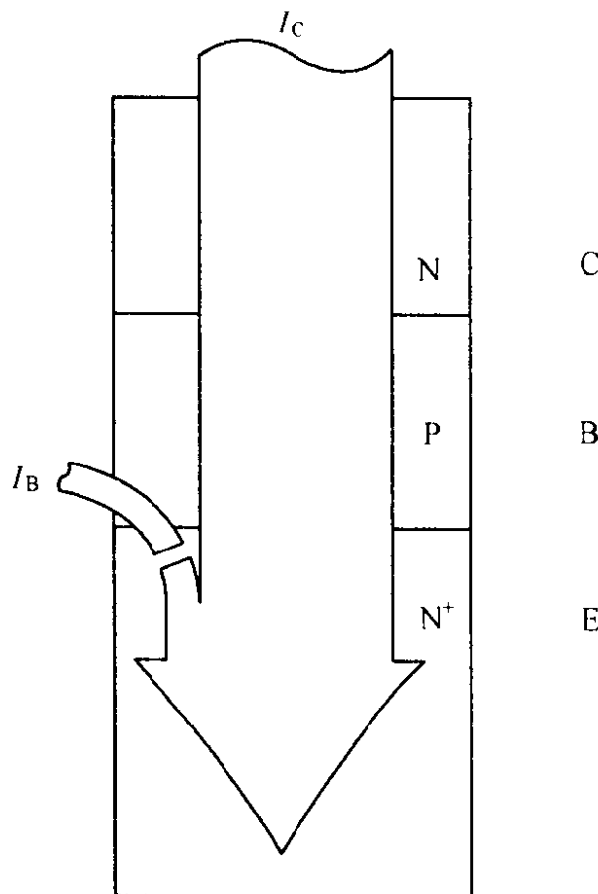


Figure 2 – 7 The terminal base and collector currents under typical conditions in a BJT

Return briefly to Figure 2 – 3(d). Focus on the carrier-density profiles in the emitter and base regions, and ignore that in the collector region because it affects only collector-leakage current, now dropped from consideration. Elementary diffusion theory permits us to write the current of electrons in the base region. This constitutes the collector current I_C , which is the output current in the common-emitter situation under study.

Having already established that it is a positive terminal current, let us employ absolute-value signs on the diffusion-current expression. Given the emitter-junction area A_E ,

$$I_C = | A_{Eq} D_{nB} (dn_B/dx) | \quad (2-2)$$

Where the gradient can be treated as constant throughout the base region, and is given simply by $n_B(0)/X_B$. It is important to stress once more at this point that X_B is taken to be the distance from boundary to boundary, and not from (metallurgical) junction to junction. It is clearly the base region as defined by the respective depletion-layer boundaries that enters into the gradient calculation. Thus X_B is the effective base thickness. To repeat, it is the thickness dimension of the active (or intrinsic) base region, whose lateral dimensions are fixed by emitter plan-view shape.

Employing the law of the junction now yields $n_B(0)$ as a function of applied voltage, in turn giving us I_C as a function of base-emitter voltage. The voltage $-V_{NP}$ in our law-of-the-junction formulation is equivalent to $-V_{EB}$, and hence to $+V_{BE}$. Therefore

$$I_C = (A_{Eq} D_{nB} n_{oB}/X_B) \exp(qV_{BE}/kT) \quad (2-3)$$

Identical principles permit us to write the current of holes diffusing into the emitter region, except that the critical length is L_{pE} rather than a device dimension. Accordingly,

$$I_B \approx | I_D | = | - A_{Eq} D_{pE} (dp_E/dx) | \quad (2-4)$$

From which

$$I_B = (A_{Eq} D_{pE} p_{oE}/L_{pE}) \exp(qV_{BE}/kT) \quad (2-5)$$

It is of the greatest importance that the same voltage V_{BE} enters into both the input and output currents. Because this is true, we can define a common-emitter static (or dc, or steady-

state) current gain that is a function of structural constants in the device:

$$\beta \equiv (I_C/I_B) \quad (2-6)$$

Making use of Equations 2-3 and 2-5 give us

$$\beta = \frac{D_{nB} n_{OB}/X_B}{D_{pE} P_{OE}/L_{pE}} \quad (2-7)$$

Thus we have from Equation 2-6

$$I_C = \beta I_B \quad (2-8)$$

Where β from Equation 2-7 the device designer and fabricator can adjust through wide ranges. Because the output current I_C is proportional to the input current I_B , the common-emitter BJT is a linear current amplifier.

The range of common-emitter current-gain values, or β values, found in practical BJTs extends through about three orders of magnitude. In power BJTs (ranging up to hundreds of watts) a combination of factors forces the designer to choose large values of X_B —several micrometers. As a result, such devices have low β —typically about ten. At the other extreme are superheat transistors, such as those at the input stages of operational amplifiers. Here the base regions are made so thin that β approaches 10 000! The price that must be paid for this kind of gain performance is a drastic loss of operating-voltage range. As the collector space-charge layer expands under its normal reverse bias, its encroachment in an ultrathin base region can permit it to reach the emitter-junction space-charge region. Under these conditions, electrons are injected directly into the collector junction so that current increases precipitously and with no further control being exerted by the input current I_B . This condition is identically the punchthrough condition treated in chapter 1.

In the era of germanium grown-junction and alloyed-junction BJTs, a large-signal model of fundamental importance was advanced by Ebers and Moll. In spite of extensive BJT structural evolution since that time chapter 1, their model remains valid. Its primary use has been for static modeling. The original paper extended the model to deal with large-signal, time-varying problems as well. But their dynamic model did not find wide acceptance, because it introduced for the extension somewhat awkward frequency-dependent current generators. For static modeling, however, the Ebers-Moll model is compact and economical. Furthermore, it is directly rooted in device physics. The physical properties stressed originally were common-base current gain, α_F and α_R , and the saturation currents exhibited by the two junctions in their BJT environment. The latter choice was a logical one in the germanium era, because as we have seen, germanium junctions are dominated by diffusion current in reverse bias and saturate cheaply. While the same approach still worked for silicon BJTs, a measure of physical relevance was lost. We now describe how that physical relevance was regained.

2.1.6 Gummel-Poon Reformulation

In 1970, a significant of the Ebers-Moll model was contributed by Gummel and Poon. It avoided the use of junction saturation currents, and common-base current gains, using instead α_F and α_R and the intercept current I_s that Gummel had previously introduced. The reformulated Ebers-Moll model contributed by Gummel and Poon has become known in its rudimentary or unelaborated embodiment as the transport form of the Ebers-Moll model. In addition to using device properties relevant to modern

silicon transistors, it is appreciably more transparent than the model in original form. For that reason, we shall develop the transport form of the model equations first and move to the original form later, showing how the two are related.

While the transport-form reformulation was a significant contribution, the Gummel-Poon paper went well beyond it. Their primary interests were (1) dynamic modeling for (2) computer analysis of BJTs, (3) incorporating realistic properties. Setting aside idealized devices, they wanted their model to deal with such items as emitter crowding, the Early effect, the Sah-Noyce-Shockley effect (space-charge-region recombination), the Webster-Rittner effect, and the Kirk effect.

Their approach to this challenge employed a variation and extension of charge-control techniques. The original methods, described and applied to the isolated junction in chapter 1, had been used extensively for the dynamic modeling of semiconductor devices. They were applied to the BJT by Beacfoy and Sparkes, with a result that is treated at some length in chapter 1. Gummel himself contributed a different kind of charge-control model and applied it to the BJT in an earlier 1970 paper. The essence of this charge-control model "of the second kind" is illustrated in Section 2.1.7. The further extension contained in the Gummel-Poon paper used a "phenomenological" or semiempirical approach to the complicated problem of the real BJT. That is, their model does not concern itself with the details of carrier distributions and physical mechanisms responsible for charge storage. For example, in the case of the Kirk effect it simply associates a certain packet of charge with the collector junction, rather than becoming involved with information such as that in Figure (which, to be

sure, was not available at that time). Their computer-oriented model employed twenty-one parameters that were to be assigned and then empirically adjusted by comparing model predictions to experimental data. It has been extensively used in the years since its introduction. The book by Getreu gives a good account of related modeling developments in the 1970s.

Finally, significant success was achieved during the late 1970s and early 1980s in generating a fully analytic high-level model for the BJT. It can be regarded as a high-level version of the Ebers-Moll model. It combines previously separate analyses of high-level junction theory and ambipolar diffusion in the BJT context. Further, it yields accurate closed-form descriptions of certain effects (especially the Kirk effect) that previously had only been treated approximately, empirically, or numerically.

2.1.7 Assumptions and Problem Definition

A highly idealized device will be specified for present purposes. To leave no doubt about the device being analyzed we offer a set of assumptions more detailed and explicit than those of Ebers and Moll:

1. The emitter, base, and collector regions individually are uniformly doped. This requires that the junctions be step junctions.
2. In the base and collector regions the doping is extrinsic but nondegenerate.
3. If the emitter-region doping is degenerate, the region can nonetheless be described by the conventional semiconductor equations by employing empirically adjusted values of quantities such as diffusivity and intrinsic density.

4. In the base and collector regions, carrier lifetime is high.
5. Injected-carrier current densities are small.
6. The I - V characteristic of each junction is of the form

$$I = - I_{K0} [\exp(-qV_{NP}/kT) - 1], \quad (2-9)$$

Where I_{K0} is a constant and is valid for that junction in the presence of the opposite junction when the opposite junction is open-circuited. (Gummel and Poon chose to employ a short-circuited condition on the opposite junction, which requires a different coefficient in the equation for the junction I - V characteristic)

7. The positional dependence of the junction boundaries upon bias voltage may be neglected.
8. Breakdown phenomena may be neglected.
9. The problem is one-dimensional.
10. The ohmic contacts are ideal and of the high-recombination-velocity variety.
11. The emitter and collector regions are thick, or extensive in the x direction, relative to their respective minority-carrier diffusion lengths.
12. The current gains α_F and α_R are independent of current and voltage.

These assumptions constitute a measure of “overkill” intended to simplify and clarify the model’s insights and construction. The Ebers-Moll model, in fact, achieved remarkable generality. In an appendix to the original paper, the authors showed that the model was practically geometry-independent. Also, we saw in section 2.1.2 that base-region doping nonuniformities have a smaller-than-expected impact upon low-level BJT properties. It was pointed out that junctions departing significantly from step-junction doping

profiles exhibit properties with great qualitative similarity to those of step junctions.

The aim of the present exercise is to write expressions for the currents at two of the BJT terminals in terms of the voltages at those terminals, with the third terminal being a common reference for defining the voltages. If current at the third terminal is of interest, it can be readily determined from the other two currents by applying Kirchhoff's current law. To define the terminal currents and voltages, it is especially convenient to use the common-base configuration; each voltage is thus applied directly to one junction only. For forward bias, the respective voltages V_{EB} and V_{CB} must obviously be negative. Ebers and Moll chose to open-circuit one junction while biasing the other; on the other hand, Gummel and Poon, as noted earlier, chose the short-circuit option. This was key step in achieving a more economical and transparent model.

2.2 The MOSFET

A thin and tightly adherent layer of glass, mainly SiO_2 , can be formed on the surface of a single-crystal silicon sample by heating it in an oxidizing atmosphere. The most common choices are oxygen and steam. Starting in the late 1950s, these procedures were moved along vigorously in the hope that they might provide at least a partial solution to the vexing problem of stabilizing surface conditions on silicon devices. Ultimately they permitted not only stabilization but also control of surface conditions; to accomplish this, however, enormous investments of time and resources were required to achieve the necessary understanding of the oxide-silicon system, and especially of the oxide-silicon interface.

At a relatively early stage, these procedures benefited the BJT, because its crucial phenomena occur mainly inside the silicon crystal. Further, even a limited grasp of oxide technology opened a whole new approach to BJT fabrication. But then came the MOSFET, which we shall describe in detail. By contrast with the BJT, the decisive phenomenon in the MOSFET occurs at the oxide-silicon interface. Consequently, the art and science of controlling interfacial properties required an order-of-magnitude improvement to make the MOSFET a truly practical device, a process that took approximately an additional ten years. The central portion of the MOSFET is an MOS capacitor, a structure that received intensive study in its own right. When isolated, it is a one-dimensional “sandwich” consisting of the oxide with silicon on one side, and with an adherent field-plate electrode on the opposite side that is capable of manipulating potential at the oxide-silicon interface.

2.2.1 Basic Mosfet Theory

As we saw in Chapter 2.1, the essence of BJT operation is applying a voltage to modulate the potential barrier in a PN junction. The modulated current of carriers spilling over the barrier as a result becomes an output current, and these combined phenomena are responsible for the remarkable gain properties of this important bipolar device. The operating principle in a field-effect transistor, or FET, is quite different. An FET is essentially a voltage-controlled resistor. Hence it is “unipolar” in the sense that carriers that are in the majority in the region providing the resistance are dominate in determining device properties. As the result of a voltage applied from one end of the resistor to the

other, these majority carriers flow ohmically in a thin region known as the channel. A separate control voltage that creates a modulated electric field essentially normal to the thin channel modulates the population of carriers in the channel. As carrier population is changed, so, field essentially normal to the thin channel. As carrier population is changed, so, too, is channel resistance. The control electrode is known as the gate electrode.

There are several FET families, each with many members, and in each family there are differences of detail in how the control voltage interacts with the channel. Nonetheless, all of the FETs have similar gain properties, properties that are substantially inferior to those of the BJT through a wide current range. But, we should add quickly, the various families have offsetting advantages. In one case, that of the MOSFET, the sum of these advantages is so imposing that MOSFETs are the dominant solid-state devices in the world today.

2.2.2 Field-Effect Transistors

Well before the present solid-state-device era, intuitive thoughts were advanced on how to realize a solid-state amplifier; the structures proposed were essentially field-effect transistors. The serious postwar effort that led to the transistor also began with field-effect transistors. While that path was being explored, an unintended experimental event led to the discovery of the point-contact transistor, which found enough applications to be characterized as the first practical solid-state amplifier. But the point-contact transistor had serious shortcomings (and operating principles that remain obscure to this day). Consequently it was supplanted by the BJT within just a few years. The events

connected with these two devices (occurring from the mid-1940s to the mid-1950s) display a fascinating blend of scientific insight, art, engineering intuition, and happenstance.

A practical FET structure was described and analyzed in detail by Shockley in 1952, a year before its reduction to practice by Foy and Wiegmann, working under the direction of Ross and Dacey. This was a case of having metaphorical lightning strike twice in the same place, because as we noted in Chapter 2.1, the same inventor had accomplished the same feat just a few years earlier with the BJT! Shockley's FET employed a pair of PN junctions for defining the channel, and hence has become known as the junction FET, or JFET. His innovation placed the region of crucial properties well inside the semiconductor crystal and away from the troublesome and poorly understood surfaces. Varying degrees of reverse bias on these channel-defining junctions caused varying depletion-layer encroachment upon the channel, thus achieving the aim of altering the areal density (number per unit area) of carriers in the thin channel.

While the laboratory JFETs reported in 1953 constituted a feasibility demonstration, they were not practical devices. The JFET requires a degree of control over areal impurity density in its channel that is about ten times better than that required, for example, in the base region of a BJT. Thus the JFET was a device awaiting new and more refined fabrication options. The first practical JFETs, made in 1960, incorporated channels produced by the exponential growth of silicon from a vapor. Today the technology of choice is usually ion implantation, which achieves unprecedented control of areal impurity density in thin layers.

In the discrete-device arena, JFETs have played a relatively

small part. Unique properties, especially low noise, have earned for them a "niche market." At first JFETs did not play a role in integrated circuits either, because their peculiar requirements found that a kind of hybrid JFET-MOSFET device could play an advantageous part in circuits dominated by the latter. Today, JFETs are being incorporated into even bipolar integrated circuits, thanks to the almost commonplace availability of ion-implantation technology. In addition, the JFET may find application in monocrystalline three-dimensional integrated circuits.

The MOSFET was the second important field-effect device to come along. The heart of its original embodiment was a metal-oxide-silicon sandwich, with initial letters responsible for the acronym MOS, as in MOSFET and MOS technology. The sandwich was in fact a parallel-plate capacitor, with one metal plate and one silicon plate. Thus the structure indeed evokes early (and unsuccessful) experiments that endeavored to modulate the majority-carrier population in a thin semiconductor layer as a way of realizing a solid-state amplifier. But important new features were present in the MOSFET, and these spelled success. First, the new device exploited the infant oxide-growth technology that started with a polished single-crystal silicon sample. Significantly, one of the MOSFET inventors was also a pioneering developer of the oxide-growth technology and student of oxide-silicon interfaces. Second, the MOSFET exploited the creation of an inversion layer in the silicon at the interface, rather than attempting to achieve a significant modulation of majority-carrier areal density in a thin sample. More specifically, the metal plate of the capacitor was used to create a potential well at the surface of the silicon, a well capable of retaining the desired carriers—

electrons in a P-doped sample or holes in an N-doped sample. This constitutes an alternative description of the inversion layer at an insulator-semiconductor interface. As an isolated entity, this inversion layer is identical to that in a grossly asymmetric junction at equilibrium. But the two cases are different in context; In the present case the carriers are confined within the potential well because they cannot penetrate the insulator on the other side of the interface, while in the PN-junction case, the inversion-layer carriers are in direct communication with the identical majority carriers on the heavily doped side of the junction. Because the MOSFET has overwhelming practical importance today, and because all field-effect devices involve similar principles, our focus in the balance of this chapter will be on the MOSFET.

The FET chronicle by no means ended with the MOSFET, however. A new family was proposed in 1966 wherein control was exercised through a metal-semiconductor junction, or Schottky junction, and reduction to practice was reported the following year. The channel carriers in this device are confined between the depletion region associated with a Schottky junction and that associated with a PN junction. (Sometimes the PN junction is replaced by a junction between an extrinsic region and a semi-insulating region) This structure was given the name MESFET, for metal-semiconductor field-effect transistor. It has been realized using a number of semiconductor materials, but GaAs is by far the most common. The MESFET, in fact, is the dominant compound-semiconductor transistor today.

A still more recent FET is the MODFET, which designates the modulation-doped FET. It creates a potential well for electrons in a region of extremely light doping, and as a result, the ionic

scattering of drifting electrons is reduced to near zero, leading to extreme values of electron mobility, well above $10^4 \text{ cm}^2/\text{v} \cdot \text{s}$ at room temperature. (Hole mobilities are typically small in the GaAs family of materials) By cooling the sample to reduce phonon activity, one can achieve electron mobilities exceeding $10^5 \text{ cm}^2/\text{V} \cdot \text{s}$. Such enhanced electron mobilities were first observed in 1978, and were realized in a field-effect transistor in 1980. This feature of the MODFET has given it a second name, HEMT, the acronym for high-electron-mobility transistor. The significance of high carrier mobility is high-speed operation of the resulting device. Still another name for the MODFET is TEGFET, for two-dimensional-electron-gas FET.

The feat of separating channel electrons in the MODFET from the impurity atoms that contribute them is accomplished by means of a heterojunction, which is an interface between two semiconductor materials of differing energy gap. Two materials often used to form a heterojunction are the binary compound GaAs and the ternary compound AlGaAs. Heterojunctions are growing in importance in semiconductor technology, making possible proliferating device innovations. One of these is yet another FET, the HIGFET, which stands for heterojunction insulated-gate field-effect transistor. Among its advantages is the ready achievement of complementary-or N-channel and P-channel-devices, which make possible superior performance.

A still further advance in the FET art uses heterojunctions to create a series of quantum wells or a superlattice, terms that describe a set of very thin, parallel layers of differing energy gap in a semiconductor crystal, usually involving binary and ternary compound semiconductors. Layer thickness must be appreciably

less than the de Broglie wavelength, and typically amounts to a few tens of angstroms. (A superlattice can also be created by doping variations, and even by light or sound waves!)

One kind of FET-like structure exploiting quantum-well layers uses them as multiple channels that are under the control of a gate electrode. Because carriers are able to move in the planes of such layers at great velocity, the resulting device is capable of high gain and fast switching. An even more recent proposal would replace the two-dimensional quantum-well layers by one-dimensional quantum-well “wires”. The transition from classical to quantum principles of operation opens new realms of device possibilities. A substantial number of variations on these and related themes already exist, and it is safe to predict that the future holds many more.

2.2.3 MOSFET Definitions

The essential structure of a MOSFET is shown in Figure 2 - 8; for purposes of explanation one can imagine that this structure had been cut out of an integrated circuit. The capacitorlike arrangement formed on and by the silicon substrate accounts for the acronym MOS, as was noted above, and as is also noted in the diagram. The metal top plate was made of aluminum in most early devices. But since about 1970, heavily doped polycrystalline silicon (polysil) has been favored. Many different insulating materials have been employed through the years, but thermally grown oxide formed in situ is currently and historically by far the most important option. This description of the silicon oxide means that the necessary silicon was supplied by the single-crystal substrate, and its growth was induced, as was noted before, by

heating the sample in an oxidizing atmosphere.

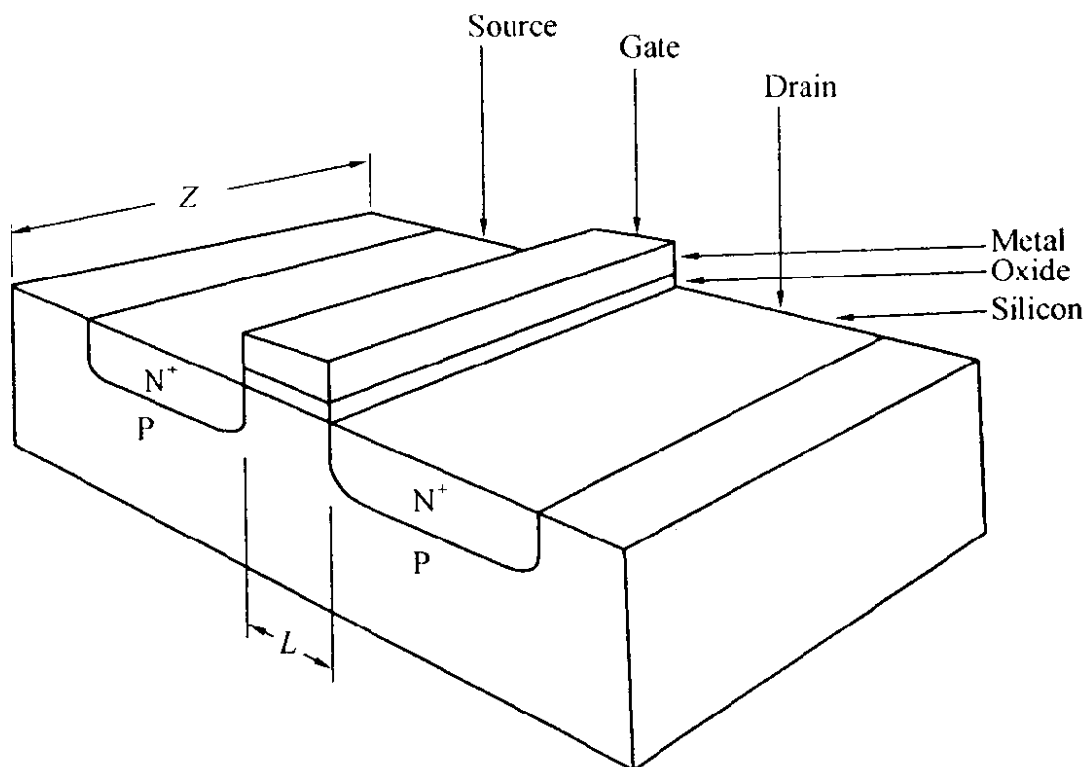


Figure 2 - 8 Essential structure of a MOSFET

The metal top plate, or field plate, becomes the control electrode of the MOSFET and is known as the gate. Applying a sufficient voltage to the gate with respect to the substrate causes electrons (normally in the minority in the P-type substrate) to be attracted to the interface, thus creating an inversion layer. This thin layer constitutes the resistor cited earlier, and is termed the MOSFET channel. Ohmic contact is made to the ends of the channel by means of heavily doped N-type regions, usually formed by ion implantation or solid-phase diffusion, or some combination of the two.

It is evident in Figure 2 - 8 that the MOSFET exhibits bilateral symmetry. That is, channel carriers can be made to flow in either direction, depending on bias polarity. This symmetry is important in certain MOSFET applications. The two channel terminals have

been named according to which one-supply carriers to the channel and which one receives carriers from it. These two terms are source and drain, respectively, and it is evident that bias polarity determines which is which. This choice of terms has been carried over from the JFET, and was made there to provide distinctive initial letters for subscript purposes. (The physicist's classically favored terms source and sink are obviously wanting with respect to this requirement.) To summarize, for the N-channel device under consideration, the more positive of the two channel terminals is the drain, the other is the source.

It is customary to refer all voltages in the MOSFET to the source voltage (just as in an earlier era, terminal voltages in a vacuum tube were referred to that of the cathode). Thus, two important bias values are the drain-source voltage V_{DS} and the gate-source voltage V_{GS} , continuing to use the double-subscript notation. But the substrate constitutes a fourth terminal. And a bias from substrate to source indeed alters device properties; this is a situation that sometimes cannot be avoided, especially in integrated circuits. To deal with it, we need at the most simplistic level a distinctive subscript for "substrate" too, and so the homely term bulk has been adopted as an approximate synonym. It is intended to connote the interior of the substrate. We shall examine the effect of a bulk source voltage V_{BS} . But for the initial description, let us assume that the source and substrate are electrically common, to serve as voltage reference. Further, for the moment let $V_{DS} = 0$.

There is a certain voltage V_{GS} at which the hole population immediately under the oxide-silicon interface will be identical to that deep within the bulk of the crystal. Under such conditions, a

band diagram representing conditions from the interface to the interior of the P-type silicon crystal would have perfectly horizontal band edges. This magic gate-bias value is termed the flat-band voltage, $V_{GS} = V_{FB}$, and is valuable as a conceptual starting point. When the gate terminal is now made more positive with respect to source (and substrate), the band edges begin to bend near the oxide-silicon interface. This matter will be examined in detail. At the moment, what is important is that potential Ψ_s at the surface of the silicon crystal (at the oxide-silicon interface) is becoming more positive through the influence of the field plate. Hole density at the interface declines and electron density rises. This is dictated by the law of mass action, $pn = n_i^2$, because the silicon remains (for engineering purposes) at equilibrium. The reason for the caveat is that any leakage current through the oxide is accompanied by a net current in the silicon that violates equilibrium. But the superb insulating qualities of SiO_2 —resistivity in excess of 10^{15} ohm $\cdot$ cm—make this a vanishing concern at present.

There exists another critical voltage at which the electron density right at the interface is equal to the whole density in the bulk. This condition is termed the threshold of strong inversion and is specified by the statement $V_{GS} = V_T$. The chosen term is apt, because channel conductance increases steeply with gate voltage for $V_{GS} > V_T$. At the threshold voltage and just below it, the electrons that form the incipient channel contribute a very small amount of drain-source conduction, leading to subthreshold current. For still smaller values of V_{GS} , there exists essentially an open circuit from drain to source because the applied voltage V_{DS} reverse biases the drain junction. Only a tiny fraction of V_{DS}

appears as forward bias on the source junction. Thus to a good approximation, one can say that the channel conductance is “turned on” by V_{GS} at the threshold voltage V_T .

2.2.4 Universal Transfer Characteristics

Treating the MOSFET simplistically as a parallel-plate capacitor with a lateral current and a resulting IR drop in one-plate neglects important considerations, as we shall see. This treatment has been accorded to two MOSFETs fabricated simultaneously, but having very different aspect ratios Z/L . Both curves have extensive linear regions. The curvature near the bottom, most evident in the higher-current device, exists because some conduction begins below the threshold voltage V_T , the phenomenon labeled subthreshold conduction. Nonetheless, because the two devices are identical in all respects except aspect ratio, extrapolating the linear portions of their I - V curves leads to a consistent and meaningful theoretical value of V_T . The value of seven volts observed for these early devices is extremely high by today's standards, and was the result of fabrication problems. A typical value today is of the order of one volt.

The device described up to this point is known as the enhancement-mode, or E-mode, MOSFET. It is designated thus because it has no channel at equilibrium; a gate voltage is required to create the inversion layer that serves as channel, and further increment of gate voltage enhance the conductivity of the channel.

It is possible, however, to create a channel that exists even in the device at equilibrium. This is usually done today by conventional doping of the region just under the gate oxide, with ion implantation being an especially favored technique. The result

is a structure somewhat like that shown in cross section in Figure 2 - 9. The channel is bounded on the bottom by a PN junction, just as in a JFET. This time a negative voltage is applied to the gate, so that the MOS capacitor requires positive charge in the silicon. Donor-ion charge fulfills that requirement. In other words, a depletion layer is formed in the N-type channel region. Once again a positive V_{DS} gives rise to an IR drop in the channel. This time it results in a thicker depletion layer at the drain end of the channel than at the source end. For the same reason, the depletion layer of the PN junction below the channel is also thicker at the drain end of the channel. Because the application of a negative voltage to the gate causes current reduction through channel depletion, this kind of device is termed a depletion-mode or D-mode MOSFET. Realized in the way just described, the D-mode MOSFET is a kind of hybrid device, with a MOSFET-like upper boundary for the channel, and a JFET-like lower boundary.

Although the channel region in the D-mode MOSFET is a bit thicker than the inversion-layer channel of the E-mode device, the D-mode MOSFET channel still qualifies as "nearly two-dimensional," and as a result, the D-mode MOSFET is also very nearly a square-law device. Its essentially parabolic current-to-voltage transfer characteristic looks like that for the E-mode device, but is translated leftward along the V_{GS} axis, as can be seen in Figure 2 - 10 (a). Thus the D-mode device has two structure-determined electrical constants: There is the characteristic current $IDSS$ at which the output current saturates when $V_{GS} = 0$. Then there is the value $V_{GS} = V_P$ that is required to "pinch off" the channel completely, so that no current can flow. Under these conditions, the depletion layer produced by gate

voltage has moved down to touch that of the PN junction, all the way from source to drain; as a result, the channel is fully depleted of carriers. The output characteristics that accompany the transfer characteristic in Figure 2 – 10(a) are shown in Figure 2 – 10(b). Only the parameter labeling of the curves distinguishes them from the E-mode characteristics.

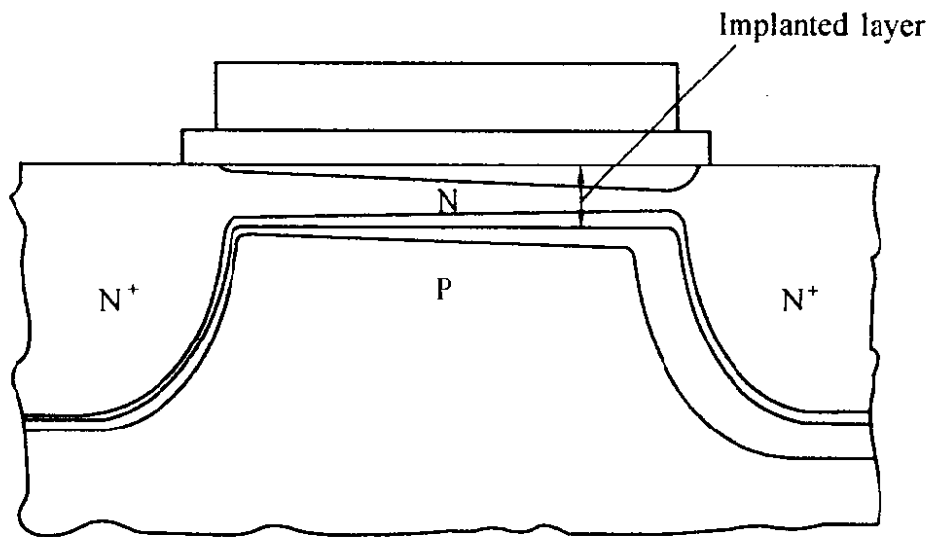


Figure 2 – 9 Schematic cross-sectional representation of a D-mode N-channel MOSFET

There is an important respect in which the D-mode MOSFET is distinguished from the JFET, with the JFET being also a D-mode device. The D-mode MOSFET possesses an enhancement regime of operation that augments its depletion regime. When positive gate bias V_{GS} is applied, additional electrons are brought into the channel from the source, and channel conductivity increases. This fact is represented in Figure 2 – 10(a) by the extension of the transfer characteristic to the right of the I_D axis, or into the regime of positive V_{GS} . The square-law behavior continues. Thus there are additional curves at higher current that can be drawn in Figure 2 – 10(b). In the JFET, by contrast, where the channel is bounded both top and bottom by PN

junctions, positive bias on the P-type gates (the top and bottom gates are usually common) will forward bias the gate junctions. Thus, for a gate bias of more than a few tenths of a volt, the gate junctions will conduct, and the high input resistance of the device

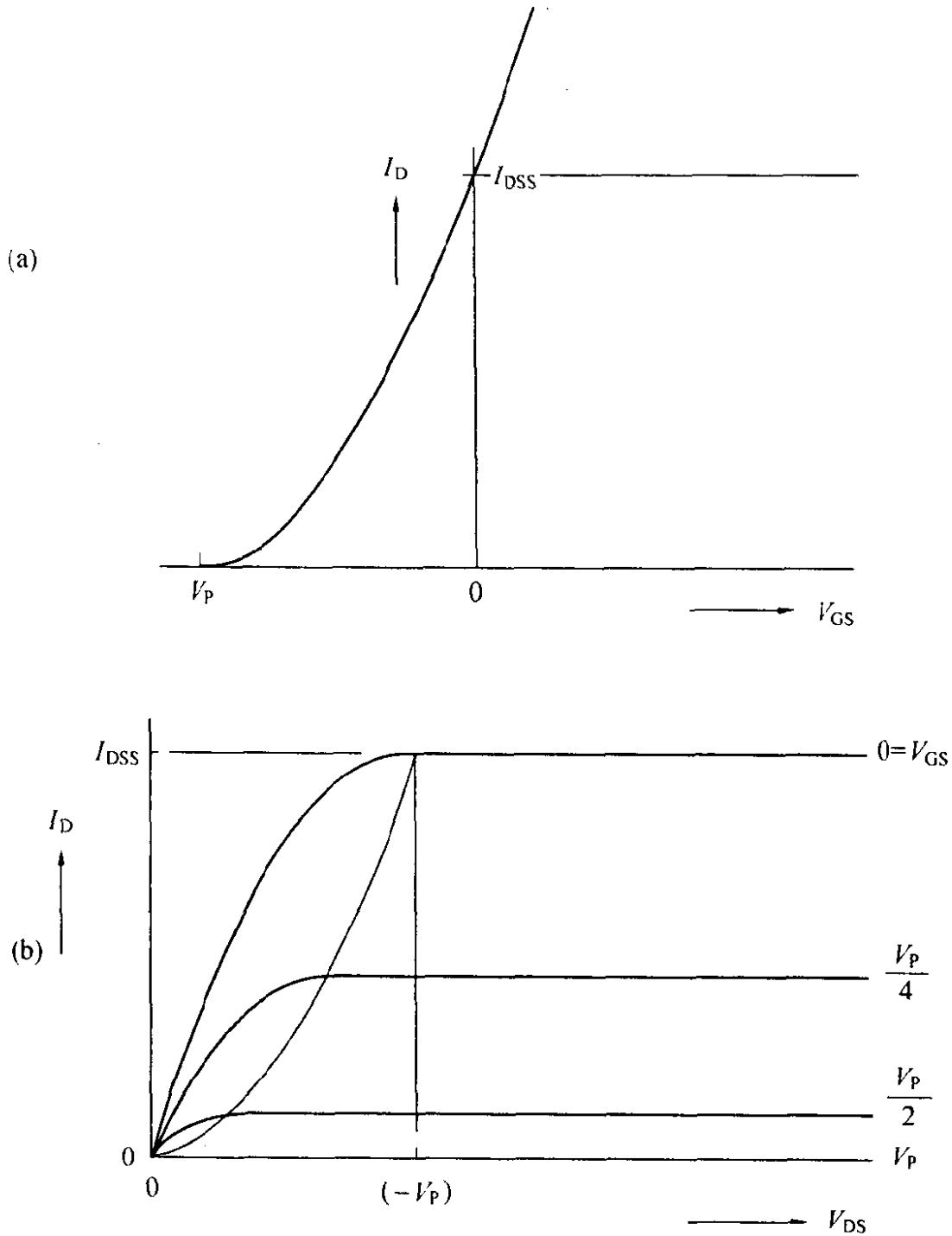


Figure 2 - 10 Current-voltage characteristics of a D-mode MOSFET

is lost, normally an undesirable situation. As a practical matter, therefore, the transfer characteristic of the N-channel JFET can be extended only a short way into the positive- V_{GS} regime that defines enhancement-mode operation.

Thus far we have considered only N-channel MOSFETs. But it is evident that by reversing all conductivity types, currents, and voltages, we can have E-mode and D-mode P-channel devices. In fact, for technological reasons P-channel devices were dominant early in the MOSFET era. It is possible to construct a transfer-characteristic diagram that summarizes all four possibilities-E and N and P. Normalization once more is convenient. The D-mode device is best treated first because it possesses a characteristic current I_{DSS} as well as a characteristic voltage V_P . Using these quantities for normalization leads to the second-quadrant curve for the D-mode N-channel MOSFET. It is simply a normalized version of the curve shown in Figure 2 - 10 (a). Since the D-mode P-channel MOSFET involves reversing currents and voltages, we merely construct a second curve in the fourth quadrant, one that is symmetric in the origin to the second-quadrant curve.

While the E-mode device does not have an explicit characteristic current for normalization, it does have an implicit characteristic current. $I_D = KV_T^2$. Using this and the voltage $V_{GS} = V_T$ for normalization, we complete Figure, displaying four curves that represent all possible MOSFETs.

Performance advantages result from combining E-mode and D-mode devices in a circuit, and such E-D products are common today. However, the combination of complementary E-mode devices, or N-channel and P-channel E-mode devices, is even more advantageous. In fact, complementary MOSFET circuits, or

CMOS circuits, constitute today ' s most rapidly growing technology.

A number of MOSFET symbols are in common use. Three informative kinds are shown in Figure 2 – 11. Distinction between a D-mode and E-mode device is sometimes maintained by using a solid line to represent the channel in the former case, and a broken line as in Figure 2 – 11(a) for the E-mode case. Another symbol, as in Figure 2 – 11(b), acknowledges that the substrate constitutes a fourth terminal of the MOSFET. Consistent with conditions in this section, we show the substrate connected (with a dotted line that is not part of the symbol) to the source. This symbol also indicates an N-channel device, with the arrow direction, as usual, from P to N, or from the P-type substrate to the channel. Finally, one sometimes wishes to denote a “preferred source,” in spite of the intrinsic symmetry of the MOSFET, and this can be done by offsetting the gate terminal in the source direction.

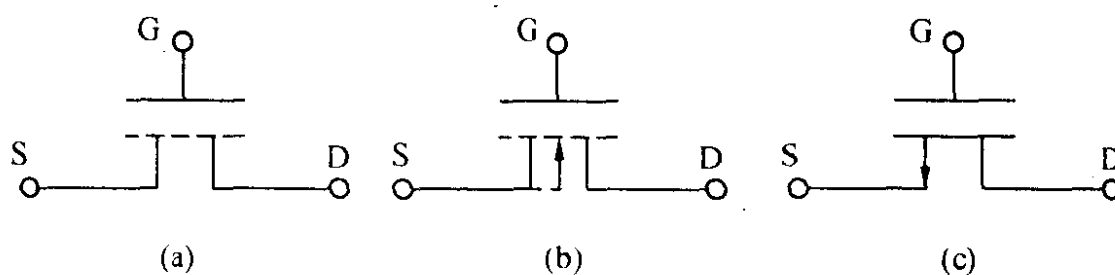


Figure 2 – 11 MOSFET symbols

A third symbol in common use is show in Figure 2 – 11(c). Its virtues are simplicity and parallelism with the BJT symbol. That is, an arrow is used to designate the source terminal and also to indicate normal current direction, just as an arrow designates the BJT emitter terminal and its normal current direction. The line connecting source and drain of course peppermints the channel,

and is sometimes made thicker to distinguish a D-mode from an E-mode device.

2.2.5 Transconductance

The rate of change of output current with respect to input voltage is an important characteristic property of a device and is termed transconductance. Its symbol is g_m . For the MOSFET,

$$g_m = \frac{\partial I_D}{\partial V_{GS}} \quad (2-10)$$

The partial derivative has conductance dimensions and has “across” or trans properties, to repeat, because it describes a result at the output port produced by a change at the input port. Transconductance exhibited in the saturation regime is of particular importance, and is of course simpler than for the curved regime.

$$g_m = 2k(V_{GS} - V_T) \quad (2-11)$$

Experimental data confirms this simple linear relationship, once again drawing data from P-channel devices of the early 1960s.

Transconductance has great importance because it plays a large part in determining the switching speed of an active device. High-transconductance devices yield circuits capable of high-speed operation. Equation 2 - 11 make it evident that MOSFET transconductance is fixed by structure and technology. That is, it depends on properties more or less under the designer's control. This explains the preference for “N-MOS” over “P-MOS” to take advantage of the higher electron mobility. It is partly responsible for the constantly declining values of X_{OX} in recent decades (thus increasing C_{OX}). Finally, the presence of the factor Z/L in g_m means that the designer makes a g_m adjustment at the graphics

console (or in an earlier era, at the drawing board).

The designer's aim, however, is to make g_m just large enough and no larger, because overdesign would involve an oversized MOSFET, requiring more silicon area. Circuit costs are directly tied to circuit area.

This brings us to another fact that helped the MOSFET toward its present dominance. Equations of simple form, and coming from the simplistic analysis just presented, give an excellent description of MOSFET properties after empirical adjustment. As a result, computer design and optimization of MOS integrated circuits has been practical since the mid-1960s. Hence the necessary fine-tuning of device sizes needed to give just enough and not too much transconductance was possible, if not easy. (It is a curious fact that the BJT is more difficult to describe with equal accuracy, but has transconductance reserves that make such description less necessary.)

As we saw, transconductance in the BJT is remarkably independent of structure and technology. It is independent of emitter area and even of the choice of semiconductor material! This is because BJT operation involves modulation of a naturally occurring potential barrier with invariant incremental properties. But in the MOSFET case, one is designing a variable resistor. Many BJT-MOSFET transconductance comparisons have been made over the years. These are not straightforward, because the two devices are so different, a subject addressed more fully in section, but they typically indicate that the BJT is superior in transconductance by a factor ranging from 10 to 100. The fact that the MOSFET is today's dominant device is the result of a large collection of other MOSFET attributes and advantages that

will be treated later. The designer, seeking always to have the best of both worlds, is today introducing the BJT into MOSFET circuits—especially CMOS circuits—in places where high transconductance is crucial, and the resulting BiCMOS products are receiving increasing attention.

2.2.6 Spice Models

There are three primary SPICE MOSFET models, identified as Levels 1, 2, and 3. Level 1 employs the rudimentary analysis of Section 2.2.1, with empirical adjustment. The most important correction is a coefficient applied to the current expressions to bring saturation-regime current into congruence with experiment and more accurate analyses. The result of such adjustment is shown in Figure, where the rudimentary model is brought into good agreement with the ionic-charge model.

The Ihantola-Moll, or ionic-charge, analysis is taken as the Level-2 SPICE mode. It is used in a form general, permitting nonzero values of bulk-source voltage. A notable departure of the Level-2 from the Level-1 analysis is a three-halves-power dependence of current on voltage, rather than the square-law dependence in the Level-1 case. Because the numerical handling of a fractional power is more time-consuming than that of an integral power, circuit designers were motivated to seek empirical modifications of the Level-2 model. In the resulting SPICE Level-3 model, they not only removed this shortcoming but also were able to deal empirically with many other effects as well, such as short-channel and channel-shortening effects. The latter is attributable to the variations of drain-junction depletion-layer thickness, and is analogous to the Early effect in the BJT. The former implies

accounting for some 3D complications, these inevitably arise when channel length approaches depletion-layer dimensions.

The semiempirical Level-1 and Level-3 models are especially advantageous for circuit analysis. But because our present interest is equally on device physics and circuit analysis, we will focus primarily on the Level-2 Ihantola-Moll model. The starting point is a generalized form for Equation that takes into account the effect of finite bulk-source voltage,

$$I_D(V_{GS}, V_{BS}, V_{DS}) = 2K \{ [V_{GS} - V_{FB}^* + 2\Psi_B - (V_{DS}/2)] V_{DS} + (2\gamma/3) [(V_{DS} - 2\Psi_B - V_{BS})^{3/2} - (-2\Psi_B - V_{BS})^{3/2}] \}, \quad (2-12)$$

where

$$K \equiv (1/2) \mu_n C_{ox} Z / L^* \quad (2-13)$$

and L^* is an elaborated version of channel length L that takes into account a technological channel-shortening effect. Impurities implanted to form the source and drain regions diffuse laterally during subsequent heat treatments by the amount L_D^* , and as a result, the channel is shortened by the amount $2 L_D^*$. Hence,

$$L^* = (L - 2L_D^*) \quad (2-14)$$

Our purpose here is to introduce the large number of possible input parameters for the Level-2 SPICE model. To compare the model with experiment, we will employ the device geometry and the current-voltage characteristics shown in Figure 2-12. Also given in the latter figure are the characteristics calculated using the Ihantola-Moll model. Figure 2-13(a) presents a static equivalent-circuit model for the ideal MOSFET described by Equation 2-12. The current generator I_D is a function of the three variables V_{GS} , V_{BS} , and V_{DS} . Figure 2-13(b) adds to the equivalent circuit the

parasitic elements present in a real MOSFET, needed for satisfactory dynamic analysis of the device.

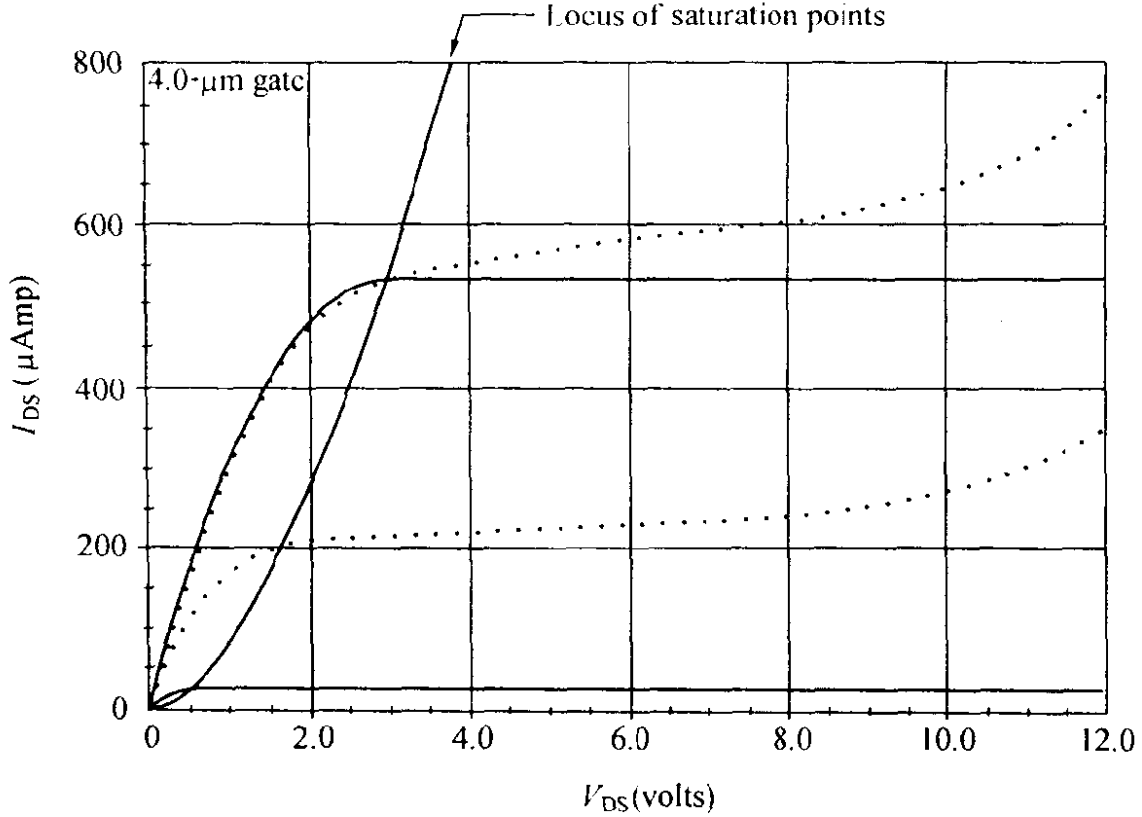


Figure 2 - 12 Three output characteristics for MOSFET

The oxide thickness X_{Ox} determines the capacitance per unit area C_{Ox} , and along with the surface mobility μ_n , enters into the current coefficient K . The substrate doping N_A and the specific capacitance C_{Ox} enter the body-effect parameter that we repeat here:

$$\gamma = \sqrt{2 \epsilon q N_A / C_{Ox}} \quad (2-15)$$

The substrate doping N_A also determines the bulk potential Ψ_{Bs} since

$$\Psi_B = (Kt/q) \ln(N_A/n_i) \quad (2-16)$$

The mobility parameters (n , UExp and ECRIT) are used to define an effective mobility as will be described later. In addition, the channel-length modulation factor, λ , is similar to the base-

thickness modulation parameter in the BJT model. The effective oxide fixed charge density N_f empirically fixes the zero-order threshold voltage determined by HD, the barrier-height difference. For metal field plates, the SPICE values agree exactly with the experimental values given by Deal and Snow, so that

$$H_D = \pm \Psi_B + 0.61 V \quad (2-17)$$

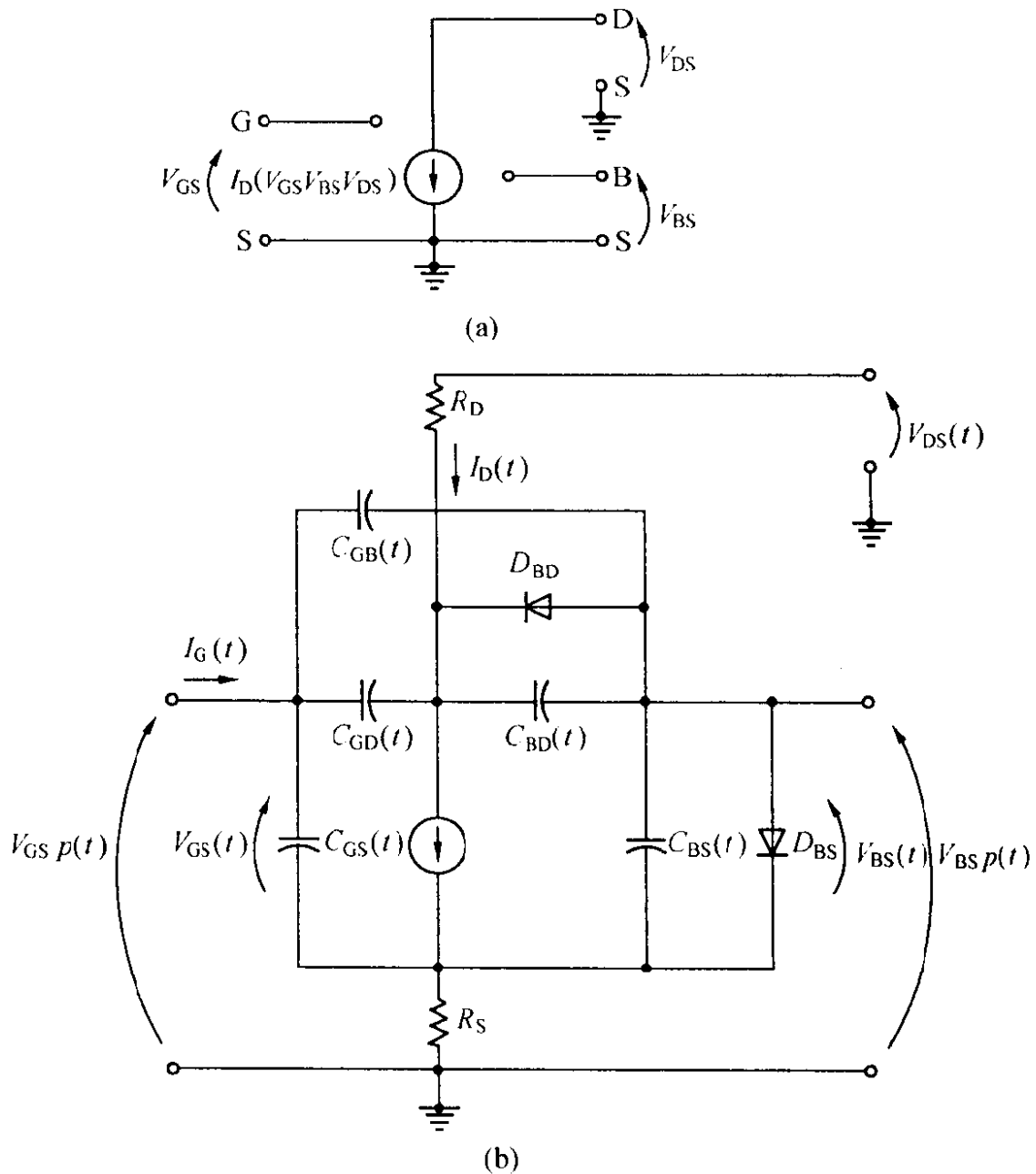


Figure 2 - 13 Level-2 modeling of the MOSFET

But for polysil field plates, we will use “effective” values N_f^* and H_D^* because the SPICE values differ slightly from the experimental results of Werner. The approximate SPICE values were chosen for convenient symmetry, with the barrier-high difference stated as

$$H_D^* = \pm [\Psi_B - (0.5575 V) T_G] \quad (2-18)$$

The positive sign represents a P-type substrate, and the negative sign, an N-type substrate. The type-of-gate parameter TG is positive for a polysil gate when the gate-doping type is opposite to that of the substrate, and negative, when the two types are the same. The effective flat-band voltage V_{FB}^* is given by

$$V_{FB}^* = H_D^* - (qN_f^* / C_{ox}) \quad (2-19)$$

And finally the threshold voltage is given by

$$V_T = V_{FB}^* - 2\Psi_B + \gamma \sqrt{-2\Psi_B} \quad (2-20)$$

To compensate for the departure of the SPICE Equation 2 - 18 from the more exact, we add a constant value of $3.3 \times 10^{10}/\text{cm}^2$, then N_f . For example, if the actual fixed-oxide charge N_f equals $3.2 \times 10^{10}/\text{cm}^2$, then N_f^* equals $6.5 \times 10^{10}/\text{cm}^2$. The internal SPICE program also employs an ancillary parameter that it terms the effective barrier height, defined as

$$V_{BI}^* = V_{FB}^* - 2\Psi_B = V_T - \gamma \sqrt{-2\Psi_B} \quad (2-21)$$

This parameter does simplify the form of some equations, but we will not use it.

Now let us return to Figure 2 - 13(b) to describe more fully the various elements it introduces. The source and drain implants have a sheet resistance R_{SH} that causes parasitic series source resistance and drain resistance, give by

$$R_S = N_{RS} R_{SH}, \quad (2-22)$$

And

$$R_D = N_{RD} R_{SH} \quad (2-23)$$

Where N_{RS} and N_{RD} are the numbers of resistive squares associated with each implanted region. The source and drain implants also produce PN junctions with a saturation-current density (extrapolated to zero bias) of J_O . The saturation currents for the corresponding junctions are given by

$$I_{BS} = A_S J_O, \quad (2-24)$$

And
$$I_{BD} = A_D J_O, \quad (2-25)$$

Where A_S and A_D are areas of the source and drain implants. In normal operation, the source and drain junctions are reverse-biased so that these diodes contribute only relatively unimportant leakage currents. However, the associated depletion-layer capacitances are important, and are divided into sidewall portions and bottom portions. The zero-bias values of the bottom parts are given by

$$C_{OBS.B} = A_S C_{O.B}, \quad (2-26)$$

And
$$C_{OBD.B} = A_D C_{O.B} \quad (2-27)$$

Where $C_{O.B}$ is the junction capacitance per unit area. The sidewall portions of the drain and source junctions contribute capacitances given by

$$C_{OBS.SW} = P_S C_{O.SW}, \quad (2-28)$$

And
$$C_{OBD.SW} = P_D C_{O.SW}, \quad (2-29)$$

Where $C_{O.SW}$ is the sidewall capacitance per unit length, P_S is the length of the sidewall for the source region, and P_D is the corresponding length for the drain region.

2.3 Microwave and Photonic Devices

2.3.1 Microwave Devices

The microwave frequencies cover the range from about 1 GHz

(10^9 Hz) to 1 000 GHz, with corresponding wavelengths from 30 to 0.03 cm. The frequencies from 30 to 300 GHz are called the millimeter wave band, because the wavelength is between 10 and 1 mm; higher frequencies are called the submillimeter wave band.

Many semiconductor devices discussed in the previous chapters can be operated in the microwave region. However, to achieve microwave capability, the device dimensions must be reduced and the parasitic capacitance and resistance must be minimized. For example, in a MESFET the gate length must be reduced to increase the cutoff frequency, and the contact resistance to the source and drain must be lowered so that the frequency response is not limited by the R_C product. Table 2 – 1 summarizes the representative microwave devices and their operational principles. In this chapter, we consider four special microwave devices: the tunnel diode, the IMPATT diode, the BARITT diode, and the transferred-electron device. We shall investigate the operational principles of these devices and how they can be used in microwave applications.

Table 2 – 1 Microwave Semiconductor Devices

Device	Operational Principle
Varactor diode	Reactance varies with bias voltage
p-i-n diode	Nearly constant capacitance, high breakdown voltage
Bipolar transistor	Electrons and holes participate in transport processes
Point contact diode	Small area, small capacitance
Schottky diode	Majority carrier transport, thermionic injection
JFET	Majority carrier, current modulated by junction gate bias

Device	Operational Principle
MESFET	Majority carrier, current modulated by Schottky gate bias
MOSFET	Majority carrier transport in surface inversion channel
Tunnel diode	Tunneling in reverse-biased $p^+ - n^+$ junction, negative differential resistance
Backward diode	Tunneling in forward-biased junction or near-zero bias
JMPATT diode	Avalanche and transit time effects to generate high power
BARITT diode	Barrier injection and transit time effects
TED	Electrons transferred from low-energy high-mobility valley to high-energy low-mobility valleys

2.3.2 Photonic Devices

Photonic devices are devices in which the basic particle of light-the photon-plays a major role. In this chapter we consider four groups of photonic devices: light-emitting diodes (LEDs) and diode lasers (light amplification by stimulated emission of radiation), which convert electrical energy to optical energy; photodetectors, which electronically detect optical signals; and solar cells, which convert optical energy into electrical energy.

The range of light detectable by the human eye extends only from approximately $0.4 \mu\text{m}$ to $0.7 \mu\text{m}$. The ultraviolet region includes wavelengths from $0.01 \mu\text{m}$ (i.e., $100 \text{ \AA}$ or 10 nm) to $0.4 \mu\text{m}$, and the infrared region extends from $0.7 \mu\text{m}$ to $1000 \mu\text{m}$. In this chapter we are primarily interested in the wavelength range from near-ultraviolet ($\sim 0.3 \mu\text{m}$) to near-infrared ($\sim 1.5 \mu\text{m}$).

To convert the wavelength to photon energy, we used the

relationship

$$\lambda = \frac{c}{\nu} = \frac{h_c}{h_\nu} = \frac{1.24}{h_\nu(\text{eV})} \mu\text{m} \quad (2-30)$$

where c is the speed of light in vacuum, ν is the light frequency, h is Planck's constant, and h_ν is the energy of a photon, and is measured in electron volts. For example, a $0.5 - \mu\text{m}$ green light corresponds to a photon energy of 2.48 eV.

There are basically three processes for interaction between a photon and an electron in a solid: absorption, spontaneous emission, and stimulated emission. We shall use a simple system to demonstrate these processes. Consider two energy levels E_1 and E_2 of an atom, where E_1 corresponds to the ground state and E_2 corresponds to an excited state (Figure 2 - 14), any transition between these states involves the emission or absorption of a photon with frequency ν_{12} given by $h\nu_{12} = E_2 - E_1$. At room temperature, most of the atoms in a solid are at the ground state. This situation is disturbed when a photon of energy exactly equal to $h\nu_{12}$ impinges on the system. An atom in state E_1 absorbs the photon and thereby goes to the excited state E_2 . The change in the energy state is the absorption process, shown in Figure 2 - 14a. The excited state of the atom is unstable; and after a short time, without any external stimulus, it makes a transition to the ground state, giving off a photon of energy $h\nu_{12}$. This process is called spontaneous emission (Figure 2 - 14b). When a photon of energy $h\nu_{12}$ impinges on an atom while it is in the excited state (Figure 2 - 14c), the atom can be stimulated to make a transition to the ground state and gives off a photon of energy $h\nu_{12}$, which is in phase with the incident radiation. This process is called stimulated emission. The radiation from stimulated emission is

monochromatic because each photon has an energy of precisely $h_{\nu_{12}}$ and is coherent because all photons emitted are in phase.

The dominant operation process for the light-emitting diode (LED) is spontaneous emission; for the laser, it is stimulated emission; and for the photo-detector and the solar cell, it is absorption.

Let us assume that the instantaneous populations of E_1 and E_2 are n_1 and n_2 , respectively. Under a thermal-equilibrium condition and for $(E_2 - E_1) > 3 kT$, the population is given by the Boltzmann distribution:

$$\frac{n_2}{n_1} = e^{-(E_2 - E_1)/kT} = e^{-h_{\nu_{12}}/kT} \quad (2 - 31)$$

The negative exponent indicates that n_2 is less than n_1 in thermal equilibrium; that is, most electrons are at the lower energy level.

In steady state, the stimulated-emission rate (i. e., the number of stimulated-emission transitions per unit time) and the spontaneous-emission rate must be balanced by the rate of absorption to maintain the population's n_1 and n_2 constant. The stimulated-emission rate is proportional to the photon field energy density $\rho(h_{\nu_{12}})$, which is the total energy in the radiation field per unit volume per unit frequency. Thus, the stimulated-emission rate can be written as $B_{21}n_2\rho(h_{\nu_{12}})$, where n_2 is the number of electrons in the upper level and B_2 is a proportionality constant. The spontaneous-emission rate is proportional only to the population of the upper level and can be written as $A_{21}n_2$ where A_{21} is a constant. The absorption rate is proportional to the electron population at the lower level and to $\rho(h_{\nu_{12}})$; this rate can be written as $B_{12}n_1\rho(h_{\nu_{12}})$, where B_{12} is a proportionality constant. Therefore, we have at steady state.

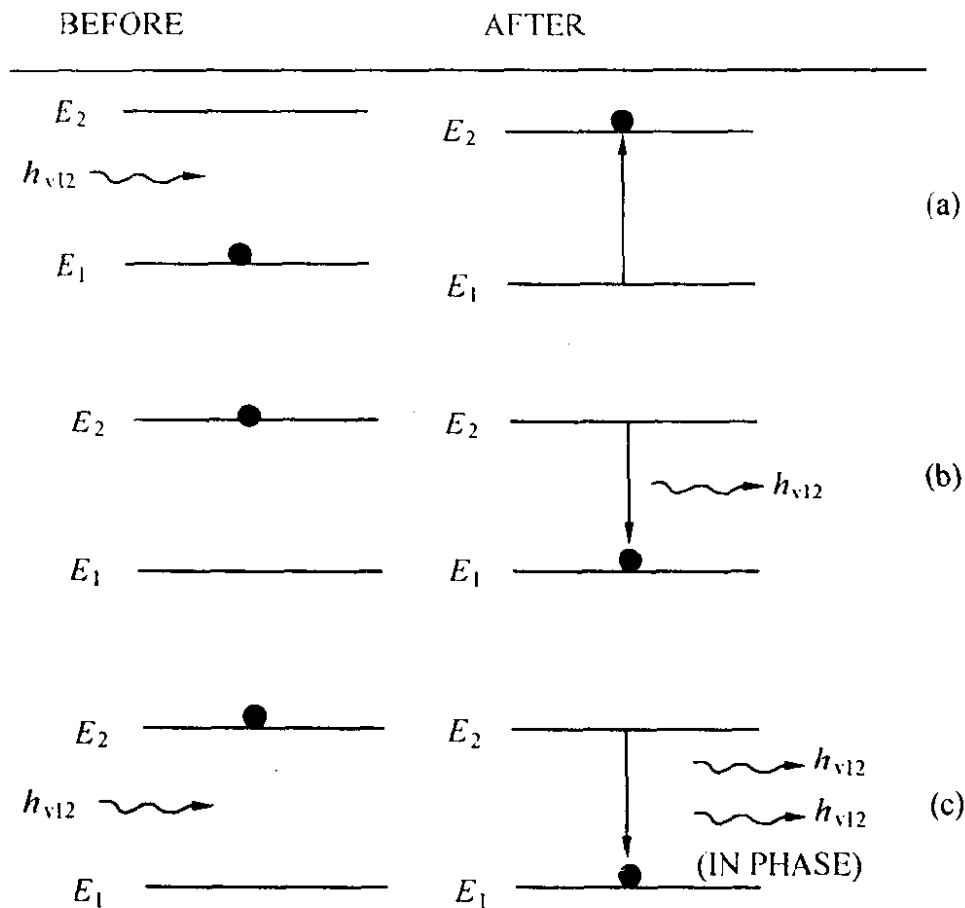


Figure 2 – 14 The three basic transition processes between two energy levels

2.3.3 Optical Absorption

Figure 2 – 15 shows the basic transitions in a semiconductor. When the semiconductor is illuminated, photons are absorbed to create electron-hole pairs as shown at (a) in Figure 2 – 15 if the photon energy is equal to the bandgap energy, that is, $h\nu$ equals E_g . If $h\nu$ is greater than E_g , an electron-hole pair is generated and, in addition, excess energy ($h\nu - E_g$) is dissipated as shown at (b) in Figure 2 – 15. Both processes (a) and (b), are called intrinsic transitions (or band-to-band transitions). On the other hand, for $h\nu$ less than E_g , a photon will be absorbed only if there

are available energy states in the forbidden bandgap due to chemical impurities or physical defects as shown at (c) in Figure 2 – 15. Process (c) is called extrinsic transition. This discussion also is generally true for the reverse situation. For example, an electron at the conduction band edge combining with a hole at the valence band edge will result in the emission of a photon with energy equal to that of the bandgap.

Assume that a semiconductor is illuminated from a light source with $h\nu$ greater than E_g and a photon flux of Φ_0 (in units of photons per square centimeter per second). As the photon flux travels through the semiconductor, the fraction of the photons absorbed is proportional to the intensity of the flux. Therefore, the number of photons absorbed within an incremental distance Δx is given by $\alpha\Phi(x)\Delta x$, where α is a proportionality constant defined as the absorption coefficient.

The fraction of photon flux that exits from the other end of the semiconductor at $x = W$ is

$$\Phi(W) = \Phi_0 e^{-\alpha x} \quad (2 - 32)$$

The absorption coefficient α is a function of $h\nu$. Figure shows the measured absorption coefficient for some important semiconductors that are used for photonic devices. Also shown is the absorption coefficient for amorphous silicon (dashed curve), which is an important material for solar cells. The absorption coefficient decreases rapidly at the cutoff wavelength λ_c ; that is,

$$\lambda_c = \frac{1.24}{E_g} \mu\text{m} \quad (2 - 33)$$

Because the optical band-to-band absorption becomes negligible for $h\nu < E_g$, or $\lambda > \lambda_c$.

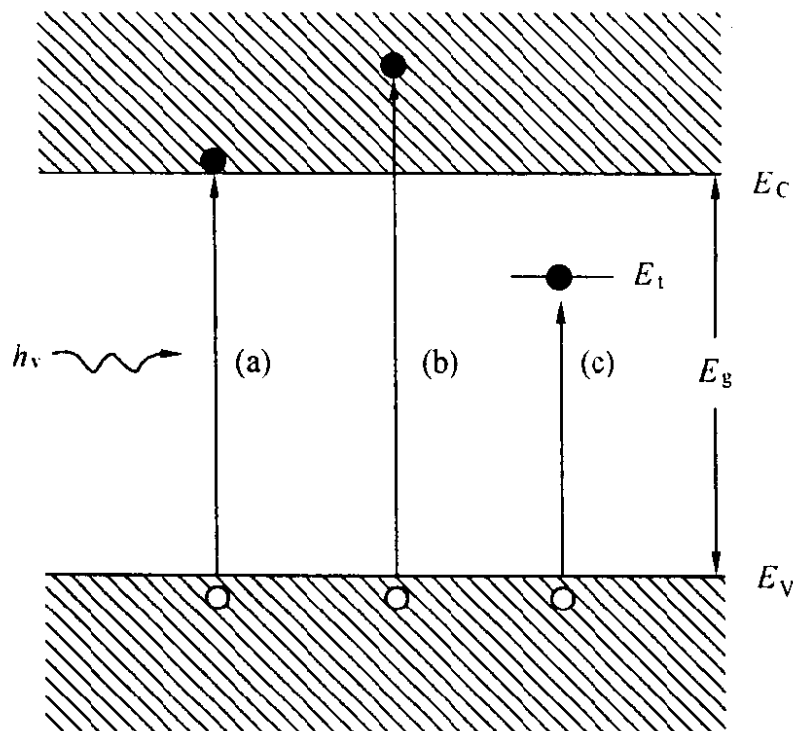


Figure 2 - 15 Optical absorption

2.4 Summary

The basic device incorporates three regions named emitter, base, and collector, each having an electrical terminal with the same name. Two PN junctions separate the three regions. Consequently, there are NPN and PNP devices. The NPN is more common and somewhat superior, but complementary circuits employing both types achieve performance and power-dissipation advantages. Conditions at the (space-charge-layer) boundaries of the two junctions are particularly important in BJT analysis. Minority-carrier densities there uniquely related to terminal voltages.

Most BJTs today are incorporated into integrated circuits. An integrated circuit is defined as a useful combination of components fabricated in (or on) a semiconductor single crystal. However, specialized BJTs such as power transistors are still being

manufactured as discrete (separately packaged) devices. The BJT is most frequently used in the common-emitter configuration, which means that the input signal is applied to the base terminal, the output signal is taken from the collector terminal, and the emitter terminal is common to the input and output ports. In this arrangement the BJT is capable of delivering both current gain and voltage gain. For normal bias in this configuration, both input (base) and output (collector) static conventional currents are positive, or inward, for the NPN device. All currents and voltages are reversed for the PNP device. The two closely spaced junctions of the BJT interact in the sense that their current values are very different from the values that would exist if the junctions were far apart and subjected to the same voltages. However, the law of the junction may still be used to find boundary values for carriers, just as though the junctions were isolated from one another. The gross asymmetry of the emitter junction causes asymmetric injection of minority carriers in its normal condition of forward bias; injection into the base region accounts for most of the junction current. The normally reverse-biased collector junction receives the injected carriers and delivers them to the collector region, where they are once more majority carriers. Transport of the minority carriers through the base region is the result of diffusion, with a nearly constant density gradient from boundary to boundary being the norm.

The area that enters in the most important way into fixing current density is emitter-junction area A_E . BJTs in common use have A_E values ranging through some eight orders of magnitude. On the other hand, the most critical thickness dimension, base thickness x_B , ranges in ordinary devices only a bit over a factor of

ten. The primary one-dimensional current through a BJT consisting of majority carriers in emitter and collector regions typically exhibits a current density that varies by only a few percent from emitter body to collector body. The density of these carriers, by contrast, typically varies by a factor of a billion or more in this same spatial interval.

The usual semiconductor approximations and assumptions are valid in the moderately doped base and collector regions that define the asymmetric collector junction. But they are not valid in the heavily doped emitter region, so empirical methods must usually be used there. The five BJT regions in the sequence of decreasing equilibrium peak-field-magnitude values are typically emitter junction, collector junction, collector region, emitter region, base region. The dominant transport mechanisms in the five regions during operation are diffusion in the emitter junction and base region and drift in the other three.

The most common BJT has emitter and base regions formed by solid-phase impurity diffusion, and hence these regions are not uniformly doped. Nonetheless, idealized step-junction analyses yield good approximations of actual BJT behavior. Although minority carriers diffusing from emitter to collector through the base region constitute a nearly one-dimensional problem, majority carriers in the base region constitute a two-or three-dimensional problem because they travel laterally from the base contact and hence into the emitter region, where nearly all of them recombine in a modern BJT. The fact that they also constitute the normally small base current I_B , however, means that their current can be neglected for some purposes. Because the base current is typically a small fraction of the total current passing through the emitter

junction, it is sometimes described as a defect current. But inasmuch as the base current is the control current in the common-emitter BJT, it plays an extremely important role. The intrinsic or active base region is defined in the thickness direction by the junction boundaries on the base sides of the two junctions and laterally by the periphery of the emitter junction (this junction usually having a smaller area than the collector junction). The balance of the base region is defined as extrinsic. In the path of the base current is the parasitic ohmic base resistance (typically 50 ohms), contributed by both the extrinsic and intrinsic base regions. Emitter and base electrical contacts having the form of closely spaced stripes tend to diminish base resistance. A small fraction of the majority carriers that constitute base current recombine in the base region with minority carriers that fail to complete the short trip from emitter junction to collector junction. In a silicon BJT, leakage current associated with the normally reverse-biased collector junction arises mainly from carrier generation in its space-charge layer.

Minority carriers diffusing through the base region constitute the collector current I_C , which is the output current in the common-emitter BJT. The ratio of I_C to the input current I_B is the static (dc) current gain β of the device. It yields to elementary analysis, except for high-level-doping effects in the emitter region that play a part in fixing I_B ; since β is determined by structural constants of the device, it is itself a constant characterizing the BJT. The relationship $I_C = \beta I_B$, stating that the BJT is a linear current amplifier, exists because both currents are fixed by emitter-junction boundary values for minority carriers, fixed in turn by a single voltage, the base-emitter voltage V_{BE} .

The input port of the common-emitter BJT is the normally forward-biased base-emitter junction, and therefore requires current biasing, or an arrangement to set static input current. Voltage division and current setting by two or more linear elements is best handled analytically. When one or more of the devices involved has a nonlinear I - V characteristic, as does forward-biased junction, a graphical technique is useful, namely the loadline diagram. It exploits the fact that (for two devices in series) the two devices carry the same current and have individual voltage drops that add to equal the applied voltage. The use of resistors for biasing in the base, collector, and (sometimes) emitter legs of a BJT circuit constitutes a practical approach to biasing. But with the advent of the integrated circuit, the use of another solid-state device (often a BJT) in lieu of one or more resistors is favored because such a device requires less silicon area than a resistor and is therefore less costly. The equivalent-circuit model constitutes a third way to handle voltage-division and current-determination problems. It is an approximate-analytic technique employing ideal devices in a configuration that yields terminal properties similar to those of the device being modeled. For the common-emitter BJT one employs a 0.7-V battery (voltage source) to model the offset voltage V_{BE} of the input-port junction, and a current-controlled current source (I_B in series with the collector terminal. The output plane is a diagram of output current I_C versus output voltage V_{CE} , with input current I_B usually employed as a parameter. For equal- I_B steps, the BJT's output characteristics in the normal-bias regime are approximately a set of equally spaced horizontal straight lines. Nonetheless, these characteristics are grossly nonlinear overall because each "turns a corner" near the I_C axis and heads

for the origin of the output plane. Consequently, the loadline construction is once again useful. When one planes in series with the collector terminal an external resistor R_C , the resistor's $I - V$ characteristic becomes the "loadline," extending from a point V_{CC} on the voltage axis, where V_{CC} is the voltage applied to R_C and the BJT, which is to say, the BJT in series, to the point V_{CC}/R_C on the current axis. The resulting diagram is useful for understanding the amplifying properties of the common-emitter BJT, which is to say, the BJT in a configuration having the emitter terminal common to the input and output ports. A positive voltage increment at the input port causes a negative increment at the output port, so the amplifier is inverting. Its small-signal voltage is fixed by component values in the biasing circuit and is proportional to current gain β .

The quiescent operation point in the common-emitter output plane of a BJT can be set by adjusting V_{CC} , R_C , and I_C (which is to say, I_B). When the emitter junction is forward-biased and the collector junction is reverse-biased, the BJT is in the forward-active regime of operation. When $V_{CE} = V_{BE}$, the collector junction has a bias voltage of zero. A still smaller value of V_{CE} means that the collector junction is forward-biased. The BJT with both junctions forward-biased is said to be in the saturation regime. The collector junction still collects efficiently, however (and collector current remains nearly constant), until its forward voltage amounts to several tenths of a volt. This is termed the regime of weak saturation. With still greater forward bias on the collector junction, collector current falls steeply, the condition known as deep saturation. Here the BJT can be modeled as two batteries- $V_{CE(SAT)} \approx 0.2 \text{ V}$ between the collector and emitter

terminals, and $V_{CE(SAT)} \approx 0.8 \text{ V}$ between the base and emitter terminals. The external circuit fixes the output-plane operating point under deep-saturation conditions.

A BJT in forward saturation has its emitter junction more forward-biased than its collector junction. Reversing the relationship places the BJT in reverse saturation. If a common-emitter BJT in the forward-active regime has bias polarities at its input and output terminals reversed, it is in the reverse-active regime of operation. When both junctions are reverse-biased (or zero-biased), the BJT is in cutoff. The simplest switching application of the BJT uses forward (deep) saturation as the ON (or closed) condition, and cutoff as the OFF (or open) condition. While a BJT in the common-emitter configuration can deliver both current gain and voltage gain, the common-base configuration can provide near-unity current gain and large voltage gain. The common-collector (or emitter-follower) configuration, on the other hand, can provide large current gain and near-unity voltage gain.

Actual BJTs exhibit parasitic properties. Examples are nonzero collector leakage current, saturation voltage, ohmic base resistance, and the Early effects. The last is a modulation of effective base thickness X_B that is caused by variations in collector-junction voltage, causing collector current and current gain (to increase with collector voltage V_{CE}). Some of the evolutionary changes in BJT structure were made to ameliorate the performance penalties associated with parasitic effects. Notable among these changes was the collector-region structure incorporating a heavily doped substrate region upon which a lightly doped layer was created by exponential growth. This structural

modification diminished the Early effect and improved breakdown voltage and collector saturation resistance.

Solid-phase-diffusion methods were introduced for creating the base and emitter-region net-doping profiles. These are attractive because of relatively easy and accurate doping control and large-batch capabilities, which translate into low cost. However, the base-region net-doping profiles. These are attractive because of relatively easy and accurate doping control and large-batch capabilities, which translate into low cost. However, the base-region net-doping profile resulting from diffusion procedures is notably nonuniform. It turns out, as Gummel demonstrated, that the details of doping profile are relatively unimportant, and that what is important is the areal density of net impurities in the base region-the net number of impurity atoms per unit area. This quantity, known as the Gummel number, ranges typically from $10^{12}/\text{cm}^2$ to $10^{13}/\text{cm}^2$.

Output characteristics (I_C versus V_{CE}) for the ideal common-emitter BJT exhibit zero conductance. In a real device, however, the Early effect, breakdown phenomena, and the punchthrough phenomenon cause departures from the ideal condition. Early-effect output conductance is approximately independent of voltage (V_{CE}) . Breakdown and punchthrough effects, however, cause output conductance to increase with voltage. Avalanche multiplication in the collector junction causes collector-emitter breakdown (with significant collector current) to occur when the multiplication factor M approaches $1/\alpha$, where α is the common-base current gain. Since α is normally close to but a bit smaller than unity, BJT breakdown voltage BV_{CEO} (base terminal open) is appreciably lower than that of an isolated junction equivalent to

the collector junction. (In the latter case, breakdown accompanies the condition wherein M approaches infinity.) In the most basic terms, the reason for this is that holes resulting from avalanche multiplication enter the base region as a "quasi base current," and this current is multiplied by the gain mechanism of the BJT. Curvature of the collector junction at its periphery has little effect on BV_{CEO} because little current passes through that portion of the junction. The negative differential resistance observable in an output characteristic for a low value of I_B occurs because α increases with current and M increases with voltage; the product of the two must equal unity for a static condition.

The ratio of forward to reverse β is significantly dependent upon the ratio of collector-junction area to emitter-junction area. Thus this BJT feature can be largely adjusted by the designer when it is important to do so. Also, possible BJT trade-offs are numerous. For example, one can achieve elimination of the Early effect and achieve a BV_{CEO} value that is essentially that of the isolate collector junction. The price of such an improvement, however, is an upper limit on collector current, and worsened (increased) saturation voltage.

The BJT is routinely used under high-level conditions, meaning that minority-carrier densities in the base region and (or) collector region approach or exceed the equilibrium majority-carrier densities there. This fact results in failure of the law of the junction, so high-level formulations of the junction problem should be used. Also, longitudinal (x -directional or junction-normal) voltage drops develop that cause imposed junction voltages (boundary-to-boundary voltages) to differ from terminal voltages. The net of these effects causes the linear curve of $\log I_C$ versus

V_{BE} to deviate sharply toward lower slope in the range $0.7 \text{ V} < V_{BE} < 0.8 \text{ V}$. The curve of $\log I_B$ versus V_{BE} , however, retains its linearity up to nearly $V_{BE} = 0.9 \text{ V}$, so that it is clear that collector current is the cause of the common-emitter current-gain (β) decline that occurs in the high-level regime. At the lower ends of these same two curves, the I_C curve retains its linearity to very low voltage, while the I_B curve deviates toward lower slope, starting in the range $0.3 \text{ V} < V_{BE} < 0.5 \text{ V}$. This is a consequence of the SNS (Sah = Noyce-Shockley) effect, and causes I_B to be responsible for the β decline that occurs in the low-level regime.

The longitudinal high-level voltage drops in the base and collector regions are accompanied by an electric field. Such a field can be resolved into an ohmic component that inevitably accompanies the passage of a net current, and a bulk-diffusion component that arises spontaneously to equalize the (apparent) diffusivities of holes and electrons in regions where density gradients exist. The base and collector regions involved remain in a quasineutral condition, for which a succinct criterion is $(dn/dx) \approx (dp/dx)$. Under such conditions, ambipolar analysis is applicable. Such analysis sometimes makes use of an (effective) ambipolar diffusivity that takes into account the net longitudinal electric field that is present, while pretending that it is absent, and an (effective) ambipolar drift mobility that takes into account the perturbation of a uniform applied electric field caused by a localized high-level condition, while pretending that the perturbations are absent. Ambipolar diffusivity in the high-level limit approaches a value intermediate between the hole and electron values. Ambipolar (drift) mobility in the high-level limit approaches zero. Drift mobility refers to the motion of a density disturbance, while

conductivity mobility refers to the net motion of individual carriers. A graded-base BJT has a base region that is nonuniformly doped, with the nonuniformities sometimes being very marked. Longitudinal electric fields, even at equilibrium, are the result of such doping variations. In graded-base devices, low-level minority-carrier profiles are very different from those in the uniformly doped case at the same current. But under high-level conditions, the doping variations become progressively less important, and nearly linear minority-carrier profiles develop for all types of variations.

The Ebers-Moll low-level large-signal model, used primarily for static analysis, is directly based upon device physics. It yields terminal current as a function of terminal voltages. It is extremely general with respect to BJT geometrical variations, but does not comprehend certain effects, such as the bias dependence of junction-boundary positions, that are sometimes important. Gummel and Poon reformulated the model (resulting in what is called the transport form) in a manner that makes it clearer and simpler for many purposes, and that adapts it more readily to the modern silicon BJT. When the original model was first presented, only germanium BJTs had reached a practical stage of development. The Ebers-Moll model exploits the linear relationship between a base-region boundary value and the current at the opposite port that exists for the common-base configuration. This inherent linearity permits the application of the superposition principle to obtain the currents existing for arbitrary boundary values and hence, via the law of the junction, for arbitrary terminal voltages. In a BJT with negligible Early effect, common-base reciprocity holds to an excellent approximation. An eight-

component equivalent-circuit Ebers-Moll model yields terminal characteristics that approximate those predicted by the two-equation analytic model.

The hybrid parameters are four partial derivatives involving terminal currents and voltages that can be treated as constant at and near a particular operation point. They are useful for linear, small-signal, dynamic characterization of the device. Because they are real numbers (and not complex numbers) they are valid only for low-frequency analysis. They lend themselves readily to equivalent-circuit application and physical interpretation. Transconductance g_m is the ratio of an incremental current delivered by an amplifying device to the incremental voltage at the input port required to produce it. In the common-emitter BJT, g_m depends on two variables only, output current I_C and inversely to T . The BJT is distinguished from competing devices by remarkably large transconductance. This is because its input voltage directly modulates the voltage barrier critical to its operation.

The low-frequency hybrid-parameter model can be extended to medium-frequency use by adding two capacitances, each associated with a junction. The resulting hybrid- π (or Giacoletto) model is useful and versatile because its elements are frequency-independent, and can be related directly to device physics. The four elements of the h -parameter model for medium-frequency use are complex numbers, and thus are distinct from those of the hybrid-parameter model, which are real. Transformation from the hybrid- π form to the h -parameter form, or the reverse, is straightforward, but sometimes tedious. The admittance model uses four complex y parameters and has advantages for analytical calculation and computer simulation. The scattering-parameter (s -

parameter) model is used for high-frequency analysis. Usually one works with the model that is most convenient for a particular purpose, and then makes transformations to other models as necessary.

A problem common to the lumped-element models is phase errors at high frequencies, even though the predictions of signal magnitudes may be adequately accurate; a distributed network is needed to model signal delay realistically. Empirical correction of phase errors is sometimes employed. A similar approach is used to introduce corrections for current-dependent base resistance, extrinsic capacitance, the Early effect, and defect current.

The charge-control model involves time constants (or frequency constants), and hence is not closely related to device physics. One starts (as in the diode case) with knowledge of stored excess-carrier charge, and then performs a spatial integration, thus giving up information on charge location. Its key assumption, that the charge distributions known for static conditions have relevance under dynamic conditions as well, is usually acceptable for small-signal cases and poor for large-signal cases. The charge-control model yields small-signal gain-versus-frequency predictions quite consistent with those of the hybrid- π and h -parameter models.

Gummel (and later, Gummel and Poon) contributed an extended and modified charge-control

model for high-level conditions. It focuses on total majority-carrier charge in the base region and writes a collector-current equation analogous to the low-level expression involving the Gummel number, but substituting the areal density of majority carriers for that of majority impurities. This approach is central in

the Gummel-Poon model.

Figures of merit are useful for evaluating and comparing devices. A useful figure of merit is independent of circuit variables and is expressible in terms of device properties. The cutoff frequency for common-emitter small-signal current gain (the frequency at which it has dropped to unity) is one example, and is useful for wide-band-amplifier applications. But because it is independent of ohmic base resistance, which plays a major role in high-frequency applications, the maximum frequency of oscillation is a more useful figure of merit for high-frequency applications.

The SPICE program for numerical analysis of the BJT can be used for both large-signal and small-signal modeling. The basic large-signal model, employing twelve equations, fourteen variables, and twenty-seven parameters, comprehends such real-life factors as parasitic series resistance, high-current effects, nonideal-diode effects, and the Early effect. From the large-signal model, one can generate hybrid-pi parameters for use in small-signal modeling. In a typical small-signal common-emitter result, one finds that Miller capacitance dominates input capacitance.

The temperature rise of a BJT in steady-state operation depends linearly upon the power dissipation involved. The constant of proportionality, known as thermal resistance, is a collective property of the materials and structures lying in the thermal path between the locus of power dissipation and the ultimate heat sink. It is measured in degrees per watt. In SPICE analysis, temperature is treated as an operating condition that affects SPICE analysis, and only steady-state temperatures are allowed.

In its early decades, the bipolar junction transistor was

known as the junction transistor or simply “the transistor.” But it in turn suffered the fate of being pushed from its position of dominance by a field-effect transistor—specifically the MOSFET treated in Chapter 2. 2. The decline of the bipolar junction transistor occurred only in relative terms, however. In absolute terms its use continues to grow.

At the heart of the MOSFET is the MOS (metal-oxide-semiconductor) capacitor, a three-layer structure with a metallic field plate on top and silicon (in the most important case) on the bottom. The intervening dielectric layer is usually SiO_2 grown thermally on the silicon. Doped regions, opposite to the substrate in type, are positioned at opposite edges of the capacitor and become two terminals of the MOSFET, with the field plate or gate being a third. The gate is the input or control electrode. In early devices it was usually made of aluminum, while today it is usually heavily doped polycrystalline silicon. Sufficient bias of the polarity that pushes majority carriers away from the oxide-silicon interface creates an inversion layer in the silicon at the interface. This is the depletion-inversion polarity. The inversion layer, or channel, permits conduction between the two-doped regions. Carriers flow from the region termed the source to the region termed the drain, with the latter having the same bias polarity as the gate. Voltages in the MOSFET are usually referred to that of the source terminal.

Majority carriers in the source and drain regions are the same as those of the channel, and these single-type flowing carriers are primarily responsible for MOSFET properties. For this reason a field-effect transistor (FET) is termed unipolar, in contrast to the bipolar junction transistor (BJT), where carriers of one type flow into the base region, and opposite-type carriers flow through

the collector junction. FETs in general can be described as voltage-controlled resistors. The MOSFET is inferior to the BJT in a number of respects. But its overall combination of properties is extremely advantageous, causing the MOSFET to be today's dominant solid-state device. MOS technology requires a degree of control of conditions in the oxide, and especially at the oxide-silicon interface, that vastly exceeds that required in BJT technology. For this reason, the MOSFET arrived about a decade later than the BJT, even though it was one of the earliest solid-state amplifiers (or switches) to be conceived in principle.

The N-channel MOSFET has an advantage over the P-channel device because electron mobility exceeds hole mobility. Inversion-layer mobilities for both, however, are typically one-half to one-third their bulk values. (The substrate region is often termed bulk for brevity and convenience.) In the N-channel case the bulk region is P-type, and we assume it to be uniformly doped (apart from the heavily and oppositely doped source and drain regions, of course). When gate-source bias voltage is adjusted to yield hole density at the oxide-silicon interface equaling that deep in the bulk region, the flatband condition exists, and the associated critical gate-source bias is termed the flat-band voltage. Variations in gate-source voltage do not disturb quasiequilibrium conditions in the substrate, so the Fermi level is constant or "flat," irrespective of band bending. When the surface potential (the mid-gap potential at the interface) equals the Fermi potential (or level), we have a condition called the threshold of weak inversion. Doubling the band bending, or causing the surface potential to be equal in magnitude and opposite in sign to the bulk potential, brings us to the threshold of strong inversion. The corresponding critical gate-

source voltage is the threshold voltage. Further increases in gate voltage cause very little further thickening of the depletion layer that preceded inversion, because it is now “easier” for MOS-capacitor lines of force to terminate on inversion-layer electrons (in the N-channel example) right at the interface than on additional negative acceptor ions at the more remote depletion-layer boundary. Band-bending character also changes qualitatively beyond threshold, with curvature increasing, because channel carriers form a thin, dense layer.

Rudimentary MOSFET analysis ignores depletion-layer charge, assuming that all charge on the silicon plate of the MOS capacitor is inversion-layer charge. It produces “square-law” (parabolic) characteristic curves in the regime of drain current versus drain-source voltage nearest the origin of the output plane. Physical reasoning predicts constant current beyond the maximum (or vertex) of each parabola, nicely confirmed experimentally in the long-channel MOSFET. This is the important saturation regime of operation, wherein drain current is independent of drain-source voltage, and proportional to the square of the gate-source voltage that is in excess of the threshold. This is the most important square-law property of the MOSFET.

The MOSFET just described is enhancement-mode (E-mode) in the sense that gate-source voltage above the threshold voltage must be applied to produce significant conductance from source to drain. By doping a thin region near the oxide-silicon interface with impurities of the same type as those in the source and drain regions, one creates a channel that exists even with the device at equilibrium. Gate-source voltage of the depletion-inversion polarity will further increase drain-source conductance. But gate-source

voltage of the opposite polarity causes a depletion layer that forms at the interface to diminish channel conductance. This is a depletion-mode (D-mode) device. The gate-source voltage that fully depletes the doped channel is known as the pinch-off voltage. Channel doping in such a device is usually accomplished by ion implantation. The D-mode MOSFET is also a square-law device.

The transconductance of the MOSFET, or the rate of change of drain current with respect to gate-source voltage, is a linear function of the latter variable. Transconductance in the MOSFET is a function of dimensional and structural properties of the device. The BJT, by contrast, has transconductance that to first order is independent of dimensional and structural properties of the device. It is proportional to output current, and hence exponentially dependent upon input voltage. As the linear-exponential comparison suggests, BJT transconductance vastly exceeds MOSFET transconductance through wide operating ranges. Transconductances in the two devices have their closest approach in the regimes of very low output current and input voltage.

An E-mode MOSFET with gate and drain common is a diode exhibiting a concave-up (square-law, again) current-voltage characteristic. Used as a load with another E-mode device as driver, it yields a linear voltage amplifier; the nonlinearities of the two devices precisely cancel. The same inverter can be used for voltage switching in a logic circuit. The diode-connected device operates always in the saturation regime, and is hence known as a saturated load. The inverter is small in area and therefore economical, but relatively slow in switching. A linear (resistive) load yields switching that is over an order of magnitude faster, but such a load is awkward technologically. A D-mode load combined

with an E-mode driver (known as an E-d inverter) is an advantageous combination, small in area, also much faster than the saturated-load inverter. One of the fastest MOS inverters N-channel and P-channel E-mode devices with gates common and drains common, the latter node serving as the output node. Such a complementary MOS (CMOS) inverter is both fastest in switching and lowest in power dissipation at low switching rates, because it requires no "standby" current; in either state of the switch, either the P-channel device is OFF, or the N-channel device is OFF.

Largely the time required to charge parasitic capacitance in the circuit fixes switching delay in MOS circuits. Thus, to optimize an integrated circuit, the sizes of its various MOSFETs (and hence their transconductances) must be adjusted according to the capacitive loads they have to "drive", so that rough parity of switching time throughout the circuit is achieved. This kind of optimization is best handled by computer, and has been so accomplished since the mid-1960s. The practice with bipolar integrated circuits has been very different because the BJT has an enormous transconductance "surplus" that to first order is geometry-independent.

The silicon portion of the MOS capacitor poses a problem that is analytically identical to the problem posed by one side of a step junction at equilibrium. Each change of bias on the capacitor, therefore, produces a condition similar to that in one region of a step junction, with doping on the other side altered. Electric field is discontinuous at the oxide-silicon interface because of the threefold ratio of dielectric constants, silicon to oxide. Treating the oxide portion as a conventional parallel-plate capacitor, and the silicon portion by means of the depletion approximation, one

can produce a serviceable model of the MOS capacitor. The potential profile in the oxide portion is linear, and in the silicon portion, parabolic. The depletion-approximation replacement (DAR) provides more accurate analysis of the silicon portion.

A real MOS capacitor embodies a number of parasitic properties. If the silicon and gate materials exhibit work-function differences, then the potential (or energy) interval from Fermi level on a given side to the conduction-band edge in the oxide will be different from that on the other side. The difference of these two potentials a term barrier-height difference. It contributes a term to threshold voltage. Another parasitic effect involves electronic states located at the surface of the silicon single crystal, into which electrons can flow when energetically favorable. That is, when band bending places these states below the Fermi level they tend to fill, and vice versa. The resulting charge in these states constitutes a highly undesirable voltage-dependent variable in the MOS capacitor. Many years of effort have devised ways to reduce the density of such "surface states" to negligible values. The carrier charges residing in such states are termed interface-trapped charge. When such states are present at high densities, they act somewhat like an electrostatic shield, because of their ability to take on and give up charge. This shielding action frustrated early field-effect experiments.

Another parasitic charge component is a quite stable positive charge near the oxide-silicon interface, believed to be associated with unreacted (unoxidized) silicon. The density of such charge correlates with silicon surface-atom density, which is a function of crystal orientation. It also correlates with the last portion of the silicon-oxidation process, because the oxidation reaction occurs at

the interface and not at the oxide surface. This oxide fixed charge, as it is known, is in contrast to the oxide mobile charge that plagued early MOS technologists. Alkali ions were mainly responsible in the latter case because of their comparatively high mobility in silicon dioxide. The cure was "cleaning up" the process, notably by avoiding the filament evaporation of a metal for the gate. The final parasitic-charge category is oxide-trapped charge. It occurs when an electron passes partway through the oxide-layer conduction band and becomes trapped, or "lodged." In an analogous way, a hole can pass part way through the oxide-layer valence band and become trapped. This can result when certain device actions and phenomena (such as avalanche breakdown) create energetic carriers near the interface. All these parasitic effects, when present, play a part in determining threshold voltage. Their combined effects were responsible for the early popularity of P-channel MOSFETs; only after control of the oxide-silicon system had advanced appreciably could the benefits of N-channel devices be exploited.

MOS-capacitor modeling and characterization are easier than step-junction modeling and characterization in the sense that the silicon remains in quasi-equilibrium in the face of bias changes. The problem is more difficult, however, in the sense that the oxide capacitance is permanently in series with the semiconductor phenomena to be modeled. The approximate-analytic method and its DAR culmination are useful for accurate capacitor modeling, especially because the MOS capacitor poses a quasiequilibrium problem.

There are several simple and useful equivalent-circuit models for the MOS capacitor, depending upon regime of operation. In

accumulation, the oxide capacitance suffices. With modest depletion, two capacitances are in series, the oxide and depletion capacitance. With the inversion layer present, as an arrangement for "feeding" carriers to it, an inversion-layer capacitance must be added to the previous model, in parallel with the depletion capacitance. With heavy inversion, we revert to the lone oxide capacitor.

The depletion and inversion capacitances are equal to each other precisely at the threshold of strong inversion. This relationship has been demonstrated and explained experimentally, theoretically, and by means of a physical model. The MOS capacitor is free of diffusion capacitance because of its insulating layer. In the junction, by contrast, diffusion capacitance and depletion capacitance are in parallel, mutually shunted by a conductance in the general case. Inversion capacitance exists in a grossly asymmetric PN junction, but it is miniscule and negative.

Observing capacitance as a function of voltage is a key method for characterizing real devices, and was the preeminent method employed for analyzing and finally understanding the oxide-silicon system. Interpretation of C-V curves is carried out in terms of the equivalent-circuit models described just above. Substrate doping, conductivity type, and oxide thickness are primary variables. The parasitic charges also affect the result, in distinctive ways in some cases. Although the formation of an inversion layer inhibits further expansion of the depletion-layer expansion by employing a voltage pulse, evening a deep-depletion characteristic. Avalanche breakdown in the depletion layer imposes a limit on deep depletion.

Transition from MOS-capacitor analysis to MOSFET analysis

means a shift from a useful and accurate one-dimensional model to one that is at best two-dimensional. For the long-channel device, a useful treatment uses the orthogonal superposition of two one-dimensional analyses.

Integrated-circuit application of any MOSFET model requires inclusion of the body effect. This term signifies that reverse biasing the bulk-source junction, a common I_C situation, means that the gate must exert a higher degree of band bending to achieve the threshold condition in the channel region than it does when that junction is not reverse-biased. In other words, the threshold voltage of such a device is increased.

Other major contributions to long-channel MOSFET modeling were made by Pao and Sah, who considered carrier diffusion in the channel and carried out numerical analyses in two orthogonal directions to achieve an extremely accurate model. Also, Baccarani et al. and Brews developed a charge-sheet model. It is relatively simple, and quite accurate as well, because of countervailing approximations.

The initial rudimentary MOSFET analysis becomes the Level-1 model in the SPICE program. Taking account of ionic charge, neglected in the first case, yields the Level-2 SPICE model. Instead of a square-law result, the ionic-charge (or Ihantola-Moll) model yields a three-halves-power result, which is awkward to treat numerically because of computation time. For this reason and others, an empirically adjusted Level-3 SPICE model was generated. It yields accurate results while avoiding the earlier shortcomings. The Level-1 model in empirically adjusted (but still square-law) form is also useful. Unadjusted, it predicts too high a current; because it assumes all charge in the silicon consists of

carriers in the channel.

SPICE considers four regimes of operation. Moving upward along a loadline, we encounter the accumulation, depletion, saturation, and curved regimes. The subthreshold regime includes portions of the first two of these. The MOSFET makes its closest approach to the remarkable BJT transconductance property in the subthreshold regime. This fact is exploited in circuitry for low-power applications, especially CMOS circuitry.

In small-signal SPICE analysis, one finds that the bulk region is approximately as effective as a control electrode as is the gate itself. The concept of current gain has meaning for the MOSFET because significant current enters the input terminal with a high-frequency signal applied and passes through parasitic capacitances. Consequently, the frequency for unity current gain, f_T , also has meaning. For the device analyzed here, it amounted to 0.5GHz.

Digital, or large-signal, applications of MOSFETs are by far the most important. The E-D combination (E-mode driver, D-mode load), is particularly advantageous, and is analyzed here. Body effect degrades the current-regulating ability of a D-mode load, and it can end up approximating a resistor. Nonetheless, the D-mode load outperforms a saturated load by more than a factor of ten in switching speed. In addition to speed, noise margin is an important feature of a digital circuit. It is a measure of the noise voltage at the input port that the circuit can absorb without changing state. Noise margin is intimately related to product reliability, yield, and economy.

Parasitic capacitances in the MOSFET that are junction-related are voltage-dependent. SPICE models this voltage dependence, and is capable of predicting the waveforms of

capacitance current, as well as resistive current. The gate-drain capacitance (Miller capacitance) is of particular importance in both digital and linear applications. Because of parasitic capacitance, the dynamic turn-on loadline lies well above the static loadline in a pulse experiment, and the turn-off loadline, well below.

A simple-theory comparison of the MOSFET and BJT with respect to transconductance versus output current shows an intersection close to but above threshold. The two curves do not actually intersect, but have their closest approach just below threshold. When input voltage is used as the independent variable, the BJT displays a single characteristic curve, having threshold voltage as a parameter. One particular comparison of normalized transconductances, BJT to MOSFET, produces an intriguing product of ratios—bulk to surface mobility, oxide to base thickness, and emitter to gate area.

When subthreshold physics is examined, the source-channel junction is seen to be an N^+N junction after the source end of the channel passes the threshold of weak inversion. This transition from NP-like to N^+N -like properties is significant, because there is an offset voltage in the conducting I-V characteristic of the former, but none in the latter. This source-end property, the electron diffusion from source to drain along linear electron profiles, and electron density at the silicon source-end surface that is exponentially dependent upon gate voltage, all combine to produce a subthreshold transconductance that approximates the BJT-like value of qI_{out}/kT .

NOTES

The individual BJTs are interconnected to perform useful circuit functions by means of a variety of methods; the result is an integrated circuit that may incorporate over 100 000 devices all fabricated simultaneously in the same silicon crystal.

分立的双极结型晶体管通过灵活互联可形成具有一定功能的实用化电路,由此可以把含有 100 000 个以上器件的电路集成在同一个硅芯片上。

The adjective “perfect” was placed in quotation marks, however, because the injected current of holes plays the critical role of the control current, and as such is a vital factor in BJT design and operation, rather than a factor to be eliminated as the potentially misleading term “defect current” may suggest.

“完美”这个形容词被置于引号之中,是由于“亏损电流”(空穴注入电流)这个术语易误解为应该消除因子,其实空穴注入电流在控制电流中起到了至关重要的作用,且这也是双极结型晶体管工作机理与设计中的重要因素。

MOS-capacitor modeling and characterization are easier than step-junction modeling and characterization in the sense that the silicon remains in quasi-equilibrium in the face of bias changes. The problem is more difficult, however, in the sense that the oxide capacitance is permanently in series with the semiconductor phenomena to be modeled.

在偏压改变时硅保持准平衡态的理论下,MOS 电容的建模和描述比突变结建模和描述要简单。然而对氧化层电容始终存在的半导体的行为描述进行建模要难得多。

Although the channel region in the D-mode MOSFET is a bit thicker than the inversion-layer channel of the E-mode device, the

D-mode MOSFET channel still qualifies as “nearly two-dimensional,” and as a result, the D-mode MOSFET is also very nearly a square-law device

尽管耗尽型 MOS 管的沟道区比增强型器件的反型层沟道稍厚,耗尽型 MOS 管沟道仍然可以近似为二维结构,由此耗尽型 MOS 管工作特性也近似为平方律关系。

New Words and Expressions

Early effects 厄尔利效应

conductance *n.* 电导

punchthrough *n.* 击穿

ambipolar *adj.* 双极的

hybrid *n.* 混合电路

transconductance *n.* 跨导

lumped-element model 集总模型

pedestal *n.* 基座;支架

orientation *n.* 定位;定向

picoampere 微微安, 10^{-12} 安培

metallurgical *adj.* 冶金学的

depletion *n.* 损耗

semiempirical *adj.* 半经验的

nondegenerate *adj.* 非简并的

polycrystalline silicon *n.* 多晶硅

filament *n.* 细丝;灯丝

shunt 分路;分路[流]器;并联

noise margin 噪声容限

metaphorical *adj.* 隐喻性的, 比喻性的

heterojunction *n.* 异质结

angstrom *n.* 埃

incipient *adj.* 初始的

saturate *v.* 饱和

congruence *n.* 相合性;一致

参 考 文 献

- 1 S M Sze. Semiconductor Devices. Bell Telephone Inc. , 1985
- 2 R M Warner, B L Grung. Semiconductor-Device Electronics. 北京: 电子工业出版社, 2002

Chapter 3

Processing Technology

3.1 Crystal Growth and Epitaxy

As discussed previously in Chapter 1, the two most important semiconductors for discrete devices and integrated circuits are silicon and gallium arsenate. In this chapter we describe the common techniques for growing single crystals of these two semiconductors. The basic process flow is from starting materials to polished wafers. The starting materials (e. g., silicon dioxide for a silicon wafer) are chemically processed to form a high-purity polycrystalline semiconductor from which single crystals are grown. The single-crystal ingots are shaped to define the diameter of the material and sawed into wafers. These wafers are etched and polished to provide smooth, specular surfaces on which devices will be made.

A technology closely related to crystal growth involves the growth of single-crystal semiconductor layers upon a single-crystal semiconductor substrate. This is called epitaxy, from the Greek words epi (meaning “on”) and taxis (meaning “arrangement”). The epitaxial process offers an important means of controlling the doping profiles so that device and circuit performances can be optimized. For example, a semiconductor layer with a relatively

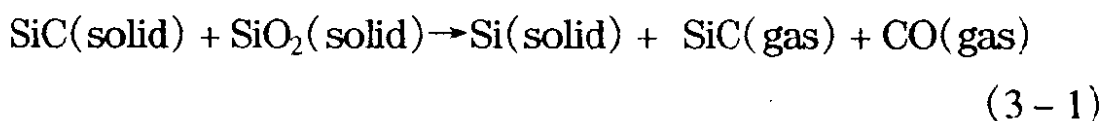
low doping concentration can be grown epitaxially upon a substrate which contains the same type of dopant in a much higher concentration (e. g. , n-type silicon on an n^+ -silicon substrate). In this way the series resistance associated with the substrate can be substantially reduced. Many novel device structures, especially for microwave and photonic devices, can be made by epitaxial processes. Later in this chapter we consider some important epitaxial growth techniques.

3.2 Crystal Growth from the Melt

There are basically two techniques for crystal growth from the melt (i. e. , material in liquid form): the Czochralski techniques and the Bridgman technique. A substantial percentage ($\sim 90\%$) of the silicon crystals for the semiconductor industry are prepared by the Csochralski technique; virtually all the silicon used for fabrication integrated circuits is prepared by this technique. Most gallium arsenide, on the other hand, is grown by the Bridgman technique. However, the Ccochralski technique is becoming more popular for the growth of large-diameter gallium arsenide.

3.2.1 Starting Materials

The starting material for silicon is a relatively pure form of sand (SiO_2) called quartzite. This is placed in a furnace with various forms of carbon (coal, coke, and wood chips). While a number of reactions take place in the furnace, the overall reaction is



This process produces metallurgical-grade silicon with a purity

of about 98%. Next, the silicon is pulverized and treated with hydrogen chloride (HCl) to form trichlorosilane (SiHCl_3):



The trichlorosilane is a liquid at room temperature (boiling point 32°C). Fractional distillation of the liquid removes the unwanted impurities. The purified SiHCl_3 is then used in a hydrogen reduction reaction to prepare the electronic-grade silicon (EGS):



This reaction takes place in a reactor containing a resistance-heated silicon rod, which serves as the nucleation point for the deposition of silicon. The EGS, a polycrystalline material of high purity, is the raw material used to prepare device quality, single-crystal silicon. Pure EGS generally has impurity concentrations in the parts-per-billion range.

The starting materials for gallium arsenate are the elemental, chemically pure gallium and arsenic, which are used for the synthesis of polycrystalline gallium arsenide. Because gallium arsenide is a combination of two materials, its behavior is quite different from that of a single material such as silicon. The behavior of a combination can be described by a phase diagram. A phase is a state (e. g., solid, liquid, or gaseous) in which a material may exist. A phase diagram shows the relationship between two components (e. g., gallium and arsenic) as a function of temperature.

Unlike silicon, which has a relatively low vapor pressure at its melting point ($\sim 10^{-6}$ atm at 1240°C), both gallium and arsenic have much higher vapor pressure at the melting point of gallium arsenide (1238°C). In its vapor phase, arsenic has As_2 and As_4 as

its major species. The vapor pressure curves for gallium arsenide melt, and the solid curves are for gallium-rich melt, more arsenic (As_2 and As_4) will be vaporized from the arsenic-rich melt, thus resulting in a higher vapor pressure. A similar argument can explain the higher vapor pressure of gallium in a gallium-rich melt. Note that long before the melting point is reached, the surface layers of liquid gallium arsenide may decompose into gallium and arsenic. Since the vapor pressure of gallium and arsenic are quite different, there is a preferential loss of the more volatile arsenic species, and the liquid becomes gallium rich.

To synthesize gallium arsenide, an evacuated, sealed quartz tube system with a two-temperature furnace is commonly used. The high-purity arsenic is placed in a graphite boat and heated to 610 to 620 °C, while the high-purity gallium is placed in another graphite boat and heated to slightly above the gallium arsenide melting temperature (1 240 to 1 260 °C). Under these conditions, an overpressure of arsenic is established (1) to cause the transport of arsenic vapor to the gallium melt, converting it into gallium arsenide, and (2) to prevent decomposition of the gallium arsenide while it is being formed in the furnace. When the melt cools, a high-purity polycrystalline gallium arsenide results. This serves as the raw material to grow single-crystal gallium arsenide.

3.2.2 The Czochralski Technique

The Czochralski technique for silicon crystal growth uses an apparatus called a puller, as shown in Figure 3 – 1. The puller has three main components: (1) a furnace, which includes a fused-silica (SiO_2) crucible, a graphite susceptor, a rotation mechanism (clockwise as shown), a heating element, and a power supply;

(2) a crystal-pulling mechanism, which includes a seed holder and a rotation mechanism (counter-clockwise); and (3) an ambient control, which includes a gas source (such as argon), a flow control, and a exhaust system. In addition, the puller has an overall microprocessor-based control system to control process parameters such as temperature, crystal diameter, pull rate, and rotation speeds, as well as to permit programmed process steps. Also, various sensors and feedback loops allow the control system to respond automatically, thereby reducing operator intervention.

In the crystal-growing process, polycrystalline silicon is placed in the crucible and the furnace is heated above the melting temperature of silicon. A suitably oriented seed crystal (e.g., $\langle 111 \rangle$) is suspended over the crucible in a seed holder. The seed is inserted into the melt. Part of it melts but the tip of the remaining seed crystal still touches the liquid surface. It is then slowly withdrawn. Progressive freezing at the solid-liquid interface yields a large, single crystal. A typical pull rate is a few millimeters per minute.

For Czochralski growth of gallium arsenide, the basic puller is identical to that for silicon. However, to prevent decomposition of the melt during crystal growth, a liquid encapsulation method is employed. The liquid encapsulate is a molten boron trioxide (B_2O_3) layer about 1 cm thick. Molten boron trioxide is inert to gallium arsenide at the growth temperature. The layer adheres to the gallium arsenide surface and serves as a cap to cover the melt. This cap prevents decomposition of the gallium arsenide as long as the pressure on its surface is higher than 1 atm (760 Torr). Since boron trioxide can dissolve silicon dioxide, the fused-silica crucible is replaced with a graphite crucible.

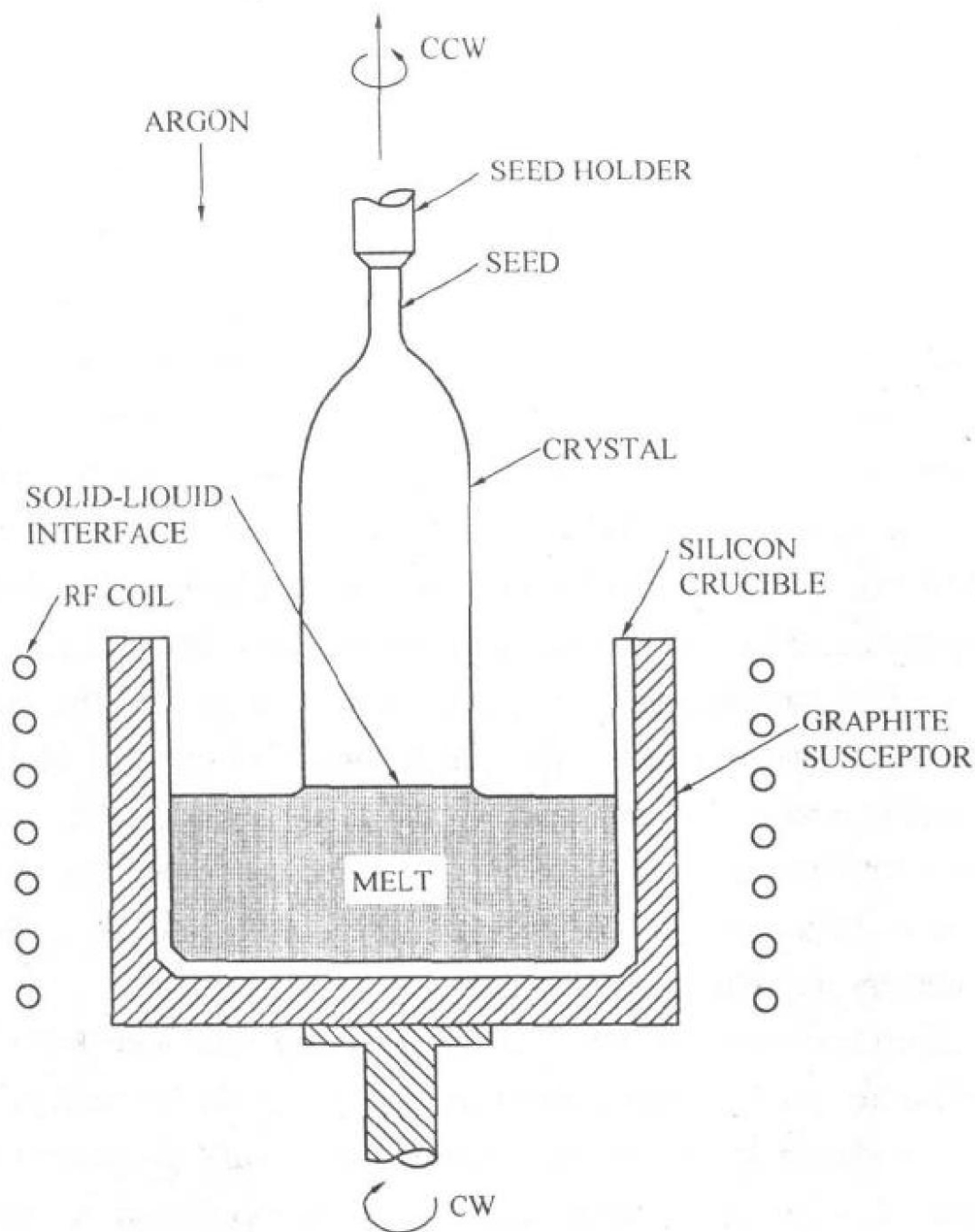


Figure 3 - 1 Czochralski crystal puller

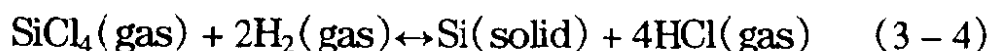
3.3 Vapor-Phase Epitaxy

In an epitaxial process, the substrate wafer acts as a seed crystal. Epitaxial processes are differentiated from the melt growth processes (described in Section 3.1) in that the epitaxial layer can

be grown at a temperature substantially below the melting point (typically 30% to 50% lower). Among various epitaxial processes, vapor phase epitaxy (VPE) is by far the most important for silicon devices. VPE is also important for gallium arsenide, but other epitaxial processes (e. g., molecular-beam epitaxy) can provide certain advantages not obtainable from VPE.

Note that the geometric shape of the susceptor provides the name for the reactor: horizontal, pancake, and barrel susceptors—all made from graphite blocks. Susceptors in the epitaxial reactors are analogous to crucibles in the crystal growing furnaces. Not only do they mechanically support the wafer, but in induction-heated reactors they also serve as the source of thermal energy for the reaction.

Four silicon sources have been used for vapor phase epitaxial growth. They are silicon tetrachloride (SiCl_4), dichlorosilane (SiH_2Cl_2), trichlorosilane (SiHCl_3), and silane (SiH_4). Silicon tetrachloride has been the most studied and has the widest industrial use. The typical reaction temperature is 1200°C . Other silicon sources are used because of lower reaction temperatures. The substitution of a hydrogen atom for each chlorine atom from silicon tetrachloride permits about a 50°C reduction in the reaction temperature. The overall reaction of silicon tetrachloride that results in the growth of silicon layers is



An additional competing reaction is taking place along with that given in Eq. 3 - 4:



As a result, if the silicon tetrachloride concentration is too high, etching rather than growth of silicon will take place. The effect of

the concentration of silicon tetrachloride in the gas on the reaction is shown, where the mole fraction is defined as the ration of the number of molecules of a given species to the total number of molecules. Note that initially the growth rate increases linearly with increasing concentration of silicon tetrachloride. As the concentration of silicon tetrachloride is increased, a maximum growth rate is reached. Beyond that, the growth rate starts to decrease, and eventually etching of the silicon will occur. Silicon is usually grown in the low concentration region.

The reaction of Eq. 3 – 4 is reversible, that is, it can take place in either direction. If the carrier gas entering the reactor contains hydrochloric acid, removal or etching will take place. Actually, this etching operation is used for in-situ cleaning of the silicon wafer prior to epitaxial growth.

The dopant is introduced at the same time as the silicon tetrachloride during epitaxial growth. Gaseous diborane (B_2H_6) is used as the p-type dopant, while phosphine (PH_3) and arsine (AsH_3) are used as n-type dopants. Gas mixtures are ordinarily used with hydrogen as the diluent to allow reasonable control of flow rates for the desired doping concentration. The dopant chemistry for arsine is illustrated in a figure, which shows arsine being adsorbed on the surface, decomposing, and being incorporated into the growing layer. To give these adsorbed atoms sufficient mobility for finding their proper positions within the crystal lattice, epitaxial growth needs relatively high temperatures

Since gallium arsenide decomposes into gallium and arsenic upon evaporation, its direct transport in the vapor phase is not possible. One approach is the use of As_4 for the arsenic component and gallium chloride ($GaCl_3$) for the gallium component. The

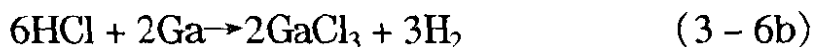
overall reaction leading to epitaxial growth of gallium arsenide is



The As_4 is generated by thermal decomposition of arsine (AsH_3):



And the gallium chloride is generated by the reaction



The reactants are introduced into a reactor with a carrier gas (e. g. , H_2). The gallium arsenide wafers are typically held within the 650 to 850 °C temperature range. There must be sufficient arsenic overpressure to prevent thermal decomposition of the substrate and the growing layer.

Another approach is the MOCVD (metalorganic chemical-vapor deposition) which uses metalorganic compounds such as trimethylgallium $\text{Ga}(\text{CH}_3)_3$. One can use trimethylgallium for the gallium component. Both chemicals can be transpired in vapor form into the reactor. The overall reaction is



During epitaxy the doping of gallium arsenide is done by the introduction of dopants in vapor form. The hydrides of sulfur and selenium or tetramethyltin are used for n-type doping; diethylzinc or diethylcadmium is used for p-type doping; and chromyl chloride is used to dope chromium onto gallium arsenide to form semi-insulation layers. As with silicon epitaxial growth, an in-situ etching is done to remove any contaminants prior to the growth.

3.4 Oxidation and Film Deposition

To fabricate discrete devices and integrated circuits we use many different kinds of thin film. We can classify thin films into

four groups: thermal oxides, dielectric layers, polycrystalline silicon, and metal films. Figure 3 – 2 shows a schematic view of a conventional silicon n-channel MOSFET that uses all four groups of film. 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 device structures. 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 the deposited 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 gate electrode material in MOS devices, as a conductive material for devices with shallow junctions. Metal films such as aluminum and silicides are used to form low-resistance interconnections, ohmic contacts to n^+ , p^+ , and polysilicon layers, and rectifying metal-semiconductor barriers.

When a film is formed (e.g., by oxidation or chemical vapor deposition), device features generally are defined by lithographic and etching processes. Each film must both perform its intended function and be compatible with the overall processing sequence, that is, the film must withstand the required chemical treatment and thermal cycle while its structure remains stable. In this chapter we consider the formation and characteristics of these films.

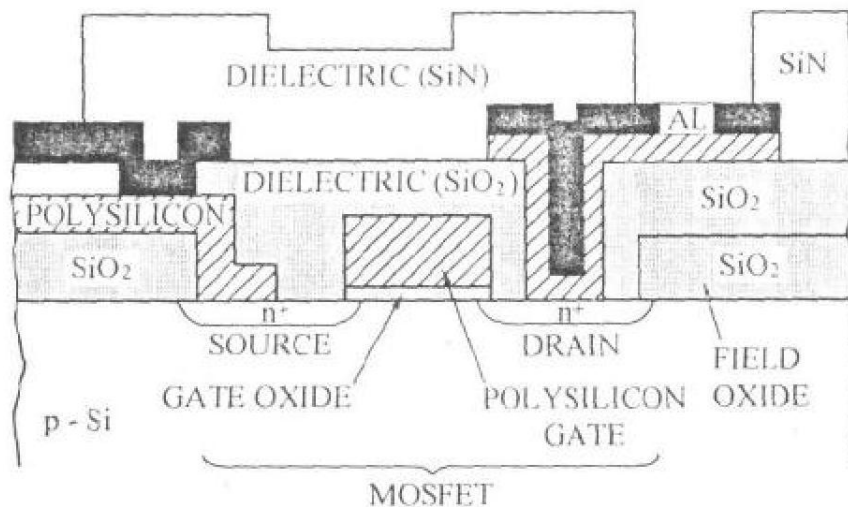


Figure 3 – 2 Schematic view of a MOSFET cross section

3.4.1 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 containing oxides of gallium and arsenic as well as arsenic ions. The oxides provide poor electrical insulation and semiconductor surface protection; hence, these oxides are rarely used in gallium arsenide technology. Consequently, in this section we shall concentrate on thermal oxidation of silicon.

The basic thermal oxidation setup is shown in Figure 3 – 3. 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 protrudes into a vertical flow hood where a filtered flow of air is maintained.

Flow is directed as shown by the arrow in Figure 3 - 3. The hood reduces dust and particulate matter in the air surrounding the wafers and minimizes contamination during wafer loading. The oxidation temperature is generally in the range of 900 to 1 200 °C, and the typical gas flow rate is about 1 cm/s. 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.

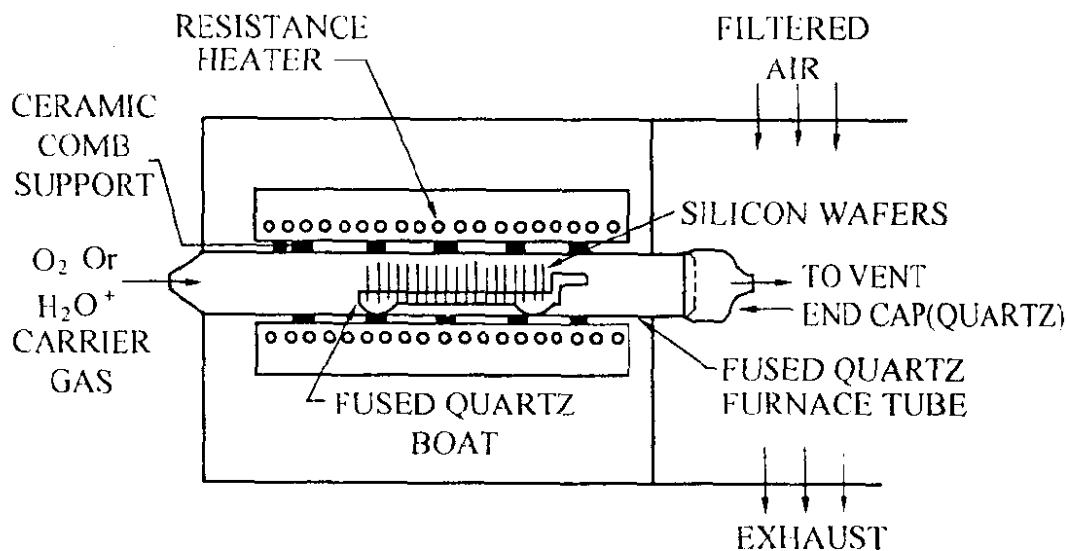
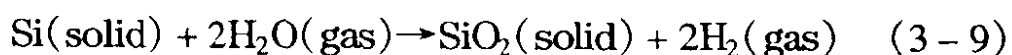
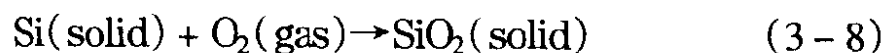


Figure 3 - 3 Schematic cross section of a resistance-heated oxidation furnace

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



The silicon-silicon dioxide interface moves into the silicon during the oxidation process. This creates a fresh interfaces 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.

3.4.2 Polysilicon Deposition

Using polysilicon as the gate electrode in MOS devices is considered a significant development in MOS circuit technology. One important reason is that polysilicon surpasses aluminum for electrode reliability. The polysilicon is clearly superior, especially for thinner gate oxides. Polysilicon is also used as diffusion source to create shallow junctions and to ensure ohmic contact to crystalline silicon. Additional uses include the manufacture of conductors and high-value resistors.

A low-pressure reactor operated between 600 and 650 °C is used to deposit polysilicon by pyrolyzing silane according to the following reaction. Of the two most common low-pressure processes, one operates at a pressure of 0.2 to 1.0 Torr using 100% silane; while the other process involves a diluted mixture of 20% to 30% silane in nitrogen at the same total pressure. Both processes can deposit polysilicon on hundreds of wafers per run with good uniformity (i.e., thickness within 5%).

At low silane partial pressure, the deposition rate is proportional to the silane pressure. At higher silane concentrations, saturation of the deposition rate occurs. Deposition at reduced pressure is generally limited to temperatures

between 600 and 650 °C. In this temperature range, the deposition rate varies as $\exp(-E_a/kT)$, where the activation energy E_a is 1.7 eV, which is essentially independent of the total pressure in the reactor. At higher temperatures, gas phase reactions that result in a rough, loosely adhering deposit become significant and silane depletion will occur, causing poor uniformity. At temperatures much lower than 600 °C, the deposition rate is too slow to be practical.

Process parameters that affect the polysilicon structure are deposition temperature, dopants, and the heat cycle applied following the deposition step. A columnar structure results when polysilicon is deposited at a temperature of 600 to 650 °C. This structure is comprised of polycrystalline grains ranging in size from 0.03 to 0.3 μm , at a preferred orientation of (110). When phosphorus is diffused at 950 °C, the structure changes to crystallite, and grain size increases to an average of between 0.5 and 1.0 μm . When temperature is increased to 1050 °C during oxidation, the grains reach a final size of 1 to 3 μm . While the initially deposited film appears amorphous when deposition occurs below 600 °C, similar growth characteristics are observed after doping and heating.

Polysilicon can be doped by diffusion, ion implantation, or the addition of dopant gases during deposition, which is referred to as in-situ doping. The implantation method is most commonly used because of its lower processing temperatures. Figure 19 shows the sheet resistance of polysilicon doped with phosphorus and antimony using ion implantation. Implant dose, annealing temperature, and annealing time all influence the sheet resistance of implanted polysilicon. Carrier traps at the grain boundaries

cause a very high resistance in the highly implanted polysilicon. Resistance drops rapidly, approaching that of implanted single-crystal silicon, as the carrier traps become saturated with dopants.

3.5 Diffusion and Ion Implantation

Diffusion and ion implantation are the two key processes we use to introduce controlled amounts of dopants into semiconductors. They are used to dope selectively the semiconductor substrate to produce either an n- or p-type region. Until the early 1970s, selective doping was done mainly by diffusion at elevated temperatures. In this method the dopant atoms are placed on or near the surface of the semiconductor 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. In this process the dopant ions are implanted into the semiconductor by means of a high-energy ion beam. The doping concentration has a peak inside the semiconductor; and the profile of the dopant distribution is determined mainly by the ion mass and the implanted-ion energy. The advantages of the ion implantation process are precise control of the total amount of dopants, improved reproducibility of impurity profiles, and lower-temperature processing.

Both diffusion and ion implantation are used for fabricating discrete devices and integrated circuits because these processes generally complement each other. For example, diffusion is used

to form a deep junction (e. g. , an n-tub in CMOS) while ion implantation is used to form a shallow junction (e. g. , a source/drain junction of a MOSFET). In this chapter we consider the basic theoretical analysis and experimental results of using these two doping methods.

3.5.1 BASIC DIFFUSION THEORY

Diffusion of impurities is typically done by placing semiconductor wafers in a furnace and passing an inert gas that contains the desired dopant through it. The furnace and gas flow arrangements are similar to those used in thermal oxidation. The temperature usually ranges between 800 and 1 200 °C for silicon and between 600 and 1 000 °C for gallium arsenide.

For diffusion in silicon, boron is the most popular dopant for introducing a p-type impurity, while arsenic and phosphorus are used extensively as n-type dopants. These three elements are highly soluble in silicon as they have solubilities above $5 \times 10^{20} \text{ cm}^{-3}$ 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). Usually, the source material is transported to the semiconductor surface by an inert carrier gas (e. g. , N_2) and is then reduced at the surface. An example of the chemical reaction for a solid source is



In this reaction an oxide layer is formed on the silicon surface.

For diffusion in gallium arsenide, because of the high vapor

pressure of arsenic special methods are used to prevent the loss of arsenic by decomposition or 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 sulfur and selenium. However, very little work has been done with this dopant.

Diffusion in a semiconductor can be visualized as atomic movement of the diffusing (dopant atoms) in the crystal lattice by vacancies or interstitials. The open circles represent the host atoms occupying the equilibrium lattice positions. The solid dots represent impurity atoms. At elevated temperatures, the lattice atoms vibrate around the equilibrium lattice sites. There is a finite probability that a host atom acquires sufficient energy to leave the lattice site and to become an interstitial atom thereby creating a vacancy. When a neighboring impurity atom migrates to the vacancy site, the mechanism is called vacancy diffusion. If an interstitial atom moves from one place to another without occupying a lattice site, the mechanism is interstitial diffusion. An atom smaller than the host atom often moves interstitially.

The basic diffusion process of impurity atoms is similar to that of charge carriers (electrons and holes) discussed in Section 1.2. Accordingly, we define a flux F as the number of dopant atoms passing through a unit area in a unit time and C as the dopant concentration per unit volume. We have

$$F = - D \frac{\partial C}{\partial x} \quad (3 - 11)$$

Where we have substituted C for the carrier concentration and the proportionality constant D is the diffusion coefficient or diffusivity. Note that the basic driving force of the diffusion process is the concentration gradient dC/dx . The flux is proportional to the concentration gradient, and the dopant atoms will move (diffuse) away from a high-concentration region toward a lower-concentration region.

3.5.2 Distribution and Range of Implanted Ions

Ion implantation is the introduction of energetic, charged particles into a substrate such as silicon. The practical use of ion implantation in semiconductor technology has been mainly to change the electrical properties of the substrate. Typical ion energies are between 30 and 300 keV, and typical ion doses vary from 10^{11} to 10^{16} ions/cm². Note that the dose is expressed as the number of ions implanted into 1 cm² of the semiconductor surface area. The main advantages of ion implantation are its more precise control and reproducibility of impurity dopings and its lower processing temperature requirements compared to those of the diffusion process.

The ion source contains the ionized dopant atoms (e.g., B⁺ and As⁺). The ions pass through a mass-separating analyzer magnet that eliminates unwanted ion species. The selected ions then enter the acceleration tube and are accelerated to high energies by an electric field. The high-energy ion beam passes through the vertical and horizontal scanners and is implanted into the semiconductor substrate.

The energetic ions lose their energy through collisions with electrons and nuclei in the substrate and finally come to rest. The

total distance which an ion travels in coming to rest is called its range R . The projection of this distance along the axis of incidence is called the projected range R_p . since the number of collisions per unit distance and the energy lost per collision are random variables, there will be a spatial distribution of ions having the same mass and the same initial energy. The statistical fluctuations in the projected range are called the projected straggle ΔR_p . There is also a statistical fluctuation along an axis perpendicular to the axis of incidence, called the lateral straggle ΔR .

Notes the ion distribution. Along the axis of incidence, the implanted impurity profile can be approximated by a Gaussian distribution function:

$$n(x) = \frac{S}{\sqrt{2\pi}\Delta R_p} \exp\left[-\frac{(x - R_p)^2}{2\Delta R_p^2}\right] \quad (3-12)$$

Where S is the ion dose per unit area. This equation is identical to Eq. 1 for constant-total-dopant diffusion, except that the quantity $4Dt$ is replaced by $2\Delta R_p^2$ and the distribution is shifted along the x -axis by R_p . Thus, for diffusion the maximum concentration is at $x = 0$, while for ion implantation the maximum concentration is at the projected range R_p . The ion concentration is reduced by 40 % from its peak value at $(x - R_p) = \pm \Delta R_p$, by one decade at $\pm 2\Delta R_p$, by two decades at $\pm 3\Delta R_p$, and by five decades at $\pm 4.8\Delta R_p$.

Along the axis perpendicular to the axis of incidence, the distribution is also a Gaussian function of the form $\exp(-y^2/2\Delta R_p^2)$. Because of this distribution, there will be some lateral implantation. However, the lateral penetration from the mask edge (of the order of ΔR) is considerable smaller than that from the thermal diffusion process discussed in Section 3.4.1.

Because of the damaged region and the disorder cluster that

result from ion implantation, semiconductor parameters such as mobility and lifetime are severely degraded. In addition, most of the ions as implanted are not located in substitutional sites. To activate the implanted ions and to restore mobility and other material parameters, we must anneal the semiconductor at an appropriate combination of time and temperature.

The annealing behaviors of boron and phosphorus implantation into silicon substrates are shown. The substrate is held at room temperature during implantation. At a given ion dose, the annealing temperature corresponds to the temperature at which 90% of the implanted ions are activated by a 30-min annealing. For boron implantation, higher annealing temperatures are needed for higher doses. Note that even at 2×10^{15} boron ions/cm², the silicon substrate remains crystalline. For phosphorus at lower doses, the annealing behavior is similar to that for boron. However, when the dose is greater than 10^{15} cm⁻², the annealing temperature drops to about 600 °C. This phenomenon is related to the solid-phase epitaxy process. At phosphorus doses greater than 10^{15} cm⁻², the silicon surface layer becomes amorphous. The single-crystal semiconductor underneath the amorphous layer serves as a seeding area for recrystallization of the amorphous layer. The epitaxial-growth rate along the $\langle 100 \rangle$ -direction is 100 Å/min at 550 °C and 500 Å/min at 600 °C, with an activation energy at 2.4 eV. Therefore, a 1 000-to 5 000-Å amorphous layer can be recrystallized in a few minutes. During the solid-phase epitaxial process, the impurity dopant atoms are incorporated into the lattice sites along with the host atoms; thus, full activation can be obtained at relatively low temperatures.

During the annealing, the implanted doping profile may be

broadened by diffusion. The initial implanted ions are given by a Gaussian distribution, Eq.3 – 12. The solution to a limited-source diffusion is also Gaussian. Therefore a solution to the diffusion equation can be given by Eq.3 – 12 with $2\Delta R_p^2$ replaced by $(2\Delta R_p^2 + 2Dt)$. For conventional furnace annealing, there is a substantial broadening of the doping profile due mainly to long annealing times (typically 30 min or longer). As device dimensions become smaller, low-temperature and short-time processing are required to minimize impurity diffusion (i. e., $\sqrt{2Dt}$ should be much less than $2\Delta R_p$). To achieve low-temperature annealing, we can implant inert ions into silicon to form an amorphous layer. The silicon can then be annealed at low temperatures ($\sim 600^\circ\text{C}$) by the solid-phase epitaxy process discussed previously to remove the implantation damage from the dopant ions.

Recently, short-time annealing processes have been studied using a variety of energy sources with a wide range of times, 10^{-9} to 10^2 s all short compared to standard furnace annealing. The e-beam annealing is done with a power density of 18 W/cm^2 for 8s after the original implant with boron (10^{15} cm^{-2} , 40 keV) and electron beam-annealed and furnace-annealed at 950°C for 30 min. The results show that e-beam annealing can activate the dopant fully with minimal redistribution, while the furnace annealing substantially broadens the profile.

Various short-time annealing techniques where the power density (in W/cm^2) is plotted against the annealing time are shown. Also indicated are the average energy densities (in parentheses). For high-power techniques (such as the pulsed laser), the silicon surface is melted, and full dopant activation is achieved via liquid-phase epitaxy growth. For lower-power

techniques (such as the CW laser), the damage is removed by the solid-phase epitaxy process. Among these techniques, the most attractive are the fast anneal technique using broad-band spectral sources (e. g., incoherent light sources) and heating with an electron beam or ion beam, because the power densities required are lower and they produce no optical-interference effects. We expect that short-time annealing will eventually replace furnace annealing, especially for very-large-scale integrated circuits where dimensional controls of dopant distributions are critical.

3.6 Lithographies

As we have discussed, the majority of IC exposure tools are optical systems using ultraviolet light. Optical exposure tools are capable of 50 to 100 wafers per hour. Electron beam lithography is primarily used to produce photomasks; relatively few are dedicated to direct wafer exposure (i. e., exposure of the resist directly by focused electron beam without a mask). Because of electron backscattering, electron beam exposure systems are limited to a practical minimum feature length of about $0.5\ \mu\text{m}$ with $0.2\ \mu\text{m}$ registration. X-Ray lithography, has $0.5\text{-}\mu\text{m}$ or better resolution and $0.5\ \mu\text{m}$ registration. However, it requires a complicated mask and is not yet used to produce ICs in volume. Ion beam lithography, offers attained-doping capability and very high resolution ($\sim 10\ \text{nm}$); this method is still in its initial development stage.

3.6.1 Electron Beam Lithography

The electron gun is a device that can generate a beam of electrons with a suitable current density; a tungsten thermionic-

emission cathode or single-crystal lanthanum boride (LaB_6) can be used for the electron gun. Condenser lenses are used to focus the electron beam to a spot size from 0.01 to $0.1\mu\text{m}$ in diameter. Beam-blanking plates (to turn the electron beam on and off) and beam deflection coils are computer-controlled and operated at high rates (in MHz or higher rates) to direct the focused electron beam to any location in the scan field on the substrate. Because the scan field (typically 1cm) is much smaller than the substrate diameter, a precision mechanical stage is used to position the substrate to be patterned.

The advantages of electron beam lithography include the generation of micron and submicron resist geometries, highly automated and precisely controlled operation, greater depth of focus that available from optical lithography, and direct patterning on a semiconductor wafer without using a mask. The disadvantage is that electron beam lithographic machines have low throughput—approximately five wafers per hour at less than $1\text{-}\mu\text{m}$ resolution. This throughput is adequate for the production of photomasks, in situations that require small numbers of custom circuits, or for design verification. However, for maskless direct writing, the machine must have the highest possible spread of the ion beam at a depth of $0.4\mu\text{m}$ is only $0.1\mu\text{m}$ in all cases [compare with Figure 16(a) for electrons]. The backscattering is completely absent for silicon substrate, and there is only a small amount of backscattering for the gold substrate.

Ion beam lithography is still in its initial development stage. Its major advantage is high spatial resolution due to absence of proximity effect. Ion beam lithography will become important when the minimum feature dimension is reduced to about $0.2\mu\text{m}$

and below.

3.6.2 Comparison of Lithographic Methods

The lithographic methods discussed above all have 1- μm or better resolution. However, each method has its limitations: diffraction effects in optical lithography, proximity effects in electron beam deflection difficulties in electron beam lithography, mask fabrication complexities in X-ray lithography, and beam deflection difficulties in ion beam lithography.

For IC fabrication, many mask levels are involved. However, it is not necessary to use the same lithographic method for all levels. A hybrid lithography approach can take advantage of the unique features of each lithographic process to improve resolution and to maximize throughput. For example, a 5:1 step-and-repeat or electron beam method can be used for the most critical mask levels, while 1:1 scanning projection can be used for the other levels.

The estimated resolution for various lithographic systems as a function of 125-mm wafer throughput is shown. For feature sizes of 1 μm or larger, the 1:1 scanning projection system is most attractive, because of its high throughput. For 1-to 0.5- μm feature sizes, we expect the M:1 step-and-repeat lithographic system with an image field of about 1 cm to dominate IC production. The throughput of electron beam systems varies as $1/l_m^2$, where l_m is the minimum feature length; this is due to the serial approach required to cover a two-dimensional area. For example, an electron beam system writing a feature size of 2 μm can have a throughput of 20 wafers per hour. The throughput for the same system drops to about one wafer per hour for a feature size of 0.5

μm . Of course, electron beam lithography will be the dominant mask-making process in the foreseeable future. When the step-and-repeat X-ray system is developed, X-ray lithography will serve as a submicron-feature-size high-throughput technique filling the gap between electron beam and optical lithography.

3.7 Wet Chemical Etching

Wet chemical etching is used extensively in semiconductor processing. Starting from the sawed semiconductor wafers, chemical etchants are used for lapping and polishing to give an optically flat, damage-free surface. Prior to thermal oxidation or epitaxial growth, the semiconductor wafers are chemically cleaned and scrubbed to remove contamination that results from handling and storing. For many discrete devices and integrated circuits of relatively large dimensions ($\geq 3 \mu\text{m}$), chemical etching is used to delineate patterns and to open windows in insulating materials. The mechanisms for wet chemical etching involve three essential steps: (1) the reactants are transported (e. g., by diffusion). Both agitation and the temperature of the etchant solution will influence the etch rate. In IC processing, most wet chemical etchings proceed by dissolution of a material in a solvent or by conversion of a material into a soluble compound which subsequently dissolves in the etching medium.

3.7.1 Isotropic Etches

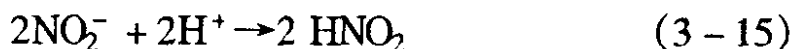
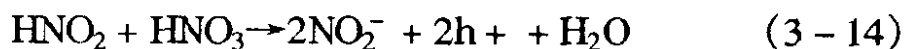
For semiconductor materials, wet chemical etching usually proceeds by oxidation, followed by the dissolution of the oxide by a chemical reaction. For silicon, the most commonly used etchants are mixtures of nitric acid (HNO_3) and hydrofluoric acid

(HF) in water or acetic acid (CH_3COOH). The reaction is initiated by promoting silicon from its initial oxidation state to some higher oxidation state and is given by



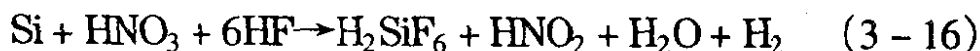
In this oxidation reaction, holes (h^+) are required. The primary oxidizing species in semiconductor etching is OH^- , which is formed by the dissociation of water:

The holes in Eq. 3 - 13 are produced by an autocatalytic process described as follows: In the presence of HNO_2 in HNO_3 solution, we have



The HNO_2 generated in the reaction, Eq. 3 - 15, reenters into reaction with HNO_3 , Eq. 3 - 14. Thus, the products of the reaction promote the reaction itself.

The overall reaction is



Water can be used as a diluent for this etchant. However, acetic acid is preferred because its use results in less dissociation of the nitric acid, hence a higher concentration of the undissociated species. Extensive studies of the HF- HNO_3 system have been made. The normally available concentrated acids are 49.2 wt% HF and 69.5 wt% of HNO_3 . We observe from these curves that at high HF and low HNO_3 concentrations, corresponding to the region near the upper vertex, the etch rate is controlled by the HNO_3 concentration (i.e., the etch rate is essentially independent of HF concentration for a given CH_3COOH diluent). This is because there is an excess of HF to dissolve the SiO_2 formed during the reaction. At low HF and high HNO_3 concentrations, corresponding to the

region near the lower right vertex, the etch rate is controlled by the ability of HF to remove the SiO₂ as it is formed; etchant solutions in this region are isotropic, that is, not sensitive to crystallographic orientation, and are used as polishing etches.

There are infinite choices of formulations for silicon etching. A number of these are well characterized. Two of the common etchants, CP-4A and CP-8. Both etchants are used for polishing of silicon wafers. Also listed are a junction staining etch to reveal the p-n junction depth and an orientation-dependent etch to be considered in the next subsection.

3.7.2 Orientation-Dependent Etching

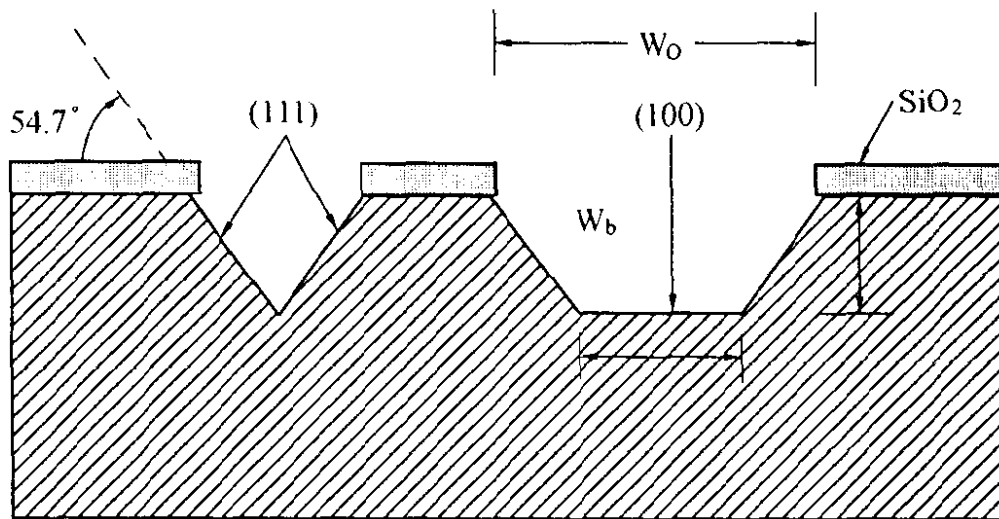
Some etchants dissolve a given crystal plane of a semiconductor much faster than other planes; this results in orientation-dependent etching. In diamond and zincblende lattices, the (111)-plane is more closely packed than the (100)-plane; therefore, the etch rate is expected to be slower for the (111)-plane. A commonly used orientation-dependent etch for silicon consists of a mixture of KOH in water and isopropyl alcohol. The etch rate is 0.6 $\mu\text{m}/\text{min}$ for the (100)-plane, 0.1 $\mu\text{m}/\text{min}$ for the (110)-plane, and only 0.006 $\mu\text{m}/\text{min}$ (6 nm/min) for the (111)-plane at about 80 °C; Thus, the ratio of the etch rates for the (100)-, (110)-, and (111)-planes is 100:16:1.

Orientation-dependent etching of $\langle 100 \rangle$ -oriented silicon through a patterned silicon dioxide mask creates precise V-shaped grooves, the edges being (111)-planes at an angle of 54.7° from the (100)-surface, as shown in the left of Figure 3 – 4(a). If the window in the mask is sufficiently large or if the etching time is short, a U-shaped groove will be formed, as shown in the right of

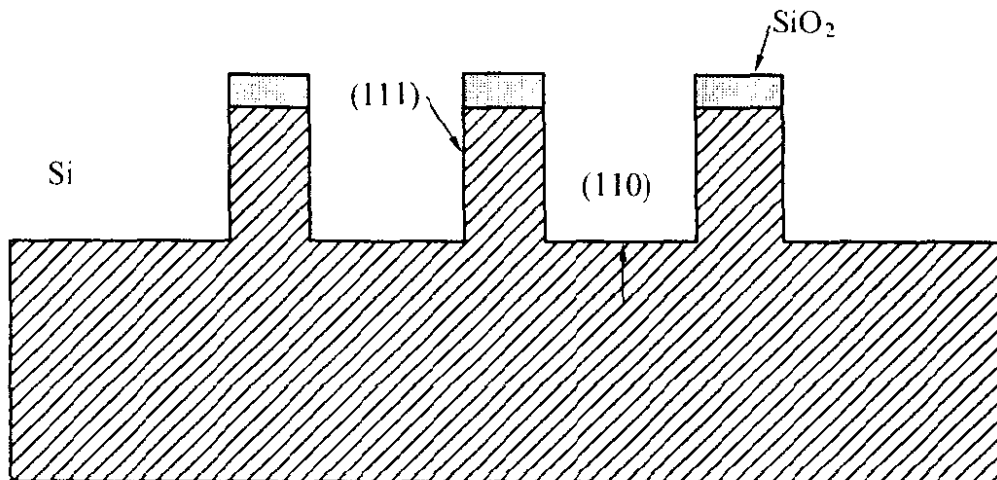
Figure 3 - 4(a). The width of the bottom surface l given by

$$W_b = W_o - 2l \cot 54.7^\circ = W_o - \sqrt{2}l \quad (3 - 17)$$

Where W_o is the width of the window on the wafer surface and l is the etched depth. If $\langle 110 \rangle$ -oriented silicon is used, essentially straight-walled grooves with sides of (111) -planes can be formed, as shown in Figure 3 - 4b. We can use the large orientation dependence in the etch rates to fabricate device structures with submicron feature lengths.



(a)



(b)

Figure 3 - 4 Orientation-dependent etching

Orientation-dependent etching in gallium arsenate is quite different from that in silicon because the (111)-gallium planes usually have the slowest etch rate, while the (111)-arsenic planes have the fastest. As a result, when the mask window is aligned with the $\langle 110 \rangle$ -axis, the etch profile is trapezoidal in one direction and dovetailed in the other. If the mask window is turned 45° with respect to the $\langle 110 \rangle$ -direction, we obtain a straight-walled groove.

3.8 Dry Etching

In pattern transfer operations, a resist pattern is defined by a lithographic process to serve as a mask for etching of its underlying layer. Most of the layer materials (e.g., SiO_2 , Si_3N_4 , and deposited metals) are amorphous or polycrystalline thin films. If they are etched in a wet chemical etchant, the etch rate is generally isotropic (i.e., the lateral and vertical etch rates are the same). Let h_f be the thickness of the layer material and l the lateral distance etched underneath the resist mask. We can define the degree of anisotropy A_f by

$$A_f = 1 - \frac{l}{h_f} = 1 - \frac{v_l t}{v_v t} = 1 - \frac{v_l}{v_v} \quad (3-18)$$

Where t is the time, v_l and v_v are the lateral and vertical etch rates, respectively. For isotropic etching, $v_l = v_v$ and $A_f = 0$.

The major disadvantage of wet chemical etching for pattern transfer is the undercutting of the layer underneath the mask, resulting in a loss of resolution in the etched pattern. In practice, for isotropic etching the film thickness should be about one third or less of the resolution required. If patterns are required with resolutions much smaller than the film thickness, anisotropic

etching (i.e., $1 \geq A_f > 0$) must be used. In practice, the value of A_f is chosen to be close to unity.

To achieve $A_f = 1$, dry etching methods have been developed. Dry etching is synonymous with plasma assisted etching, which denotes several techniques that use plasma in the form of low-pressure discharges. These techniques are commonly used in very-large-scale IC processing because of their capability for high-fidelity transfer of the resist patterns.

3.8.1 Plasma-Assisted Etching Techniques

A plasma is a fully or partially ionized gas composed of ions, electrons, and neutrons. A plasma is produced when an electric field of sufficient magnitude is applied to a gas, causing the gas to break down and become ionized. The plasma is initiated by free electrons that are released by some means such as field emission from a negatively biased electrode. The free electrons gain kinetic energy from the electric field. In the course of their travel through the gas, the electrons collide with gas molecules and lose their energy. The energy transferred in the collisions causes the gas molecules to be ionized (i.e., to free electrons). The freed electrons gain kinetic energy from the field, and the process continues. Therefore, when the applied voltage is larger than the breakdown potential, a sustained plasma is formed throughout the reaction chamber.

The electron concentrations in the plasma for dry etchings are relatively low, typically on the order of 10^9 to 10^{12} cm^{-3} . At a pressure of Torr, the concentrations of gas molecules are 10^4 to 10^7 times higher than the electron concentrations. This results in an average gas temperature in the range 50 to 100 $^{\circ}\text{C}$. Therefore, the

plasma-assisted dry etching is a low-temperature process.

Figure 3 – 5 shows two dry-etching systems. Figure 3 – 5(a) is a sputter etching system, which uses relatively high-energy (≥ 500 eV) noble gas ions, such as argon ions (Ar^+). The wafer to be etched (also called the target) is placed on the powered electrode, and argon ions are accelerated by the applied field to bombard the target surface. Through the transfer of a momentum, atoms near the surface become volatile and are removed (or etched). The typical operating pressure for sputter etching is 0.01 to 0.1 Torr. The direction of the electric field is normal to the target surface so that under the operating pressure, argon ions arrive predominantly normal to the surface. There are essentially no ion bombardments on the sidewalls of the etched features. As a result, the lateral etch rate V_v and a high degree of anisotropy can be obtained. However, there is one major drawback of the sputter-etching process: poor selectivity, that is, the etch rates for most materials are very close, and we cannot etch only one layer and stop at the underlying material.

Figure 3 – 5(b) shows a parallel-plate plasma-etching system. The plasma is confined between the two closely spaced electrodes. Molecular gases containing one or more halogen atoms (e. g., CCl_4 and Cl_2) are fed through the gas ring. The typical operating pressure is relatively high from 0.1 to 10 Torr. Another plasma-assisted etching method is reactive-ion etching (RIE), which employs apparatus similar to that for sputter etching (Figure 3 – 5a). However, in RIE the noble-gas plasma is replaced by molecular-gas plasma similar to that in plasma etching. Under appropriate conditions, both RIE and plasma etching can give high selectivity and a high degree of anisotropy.

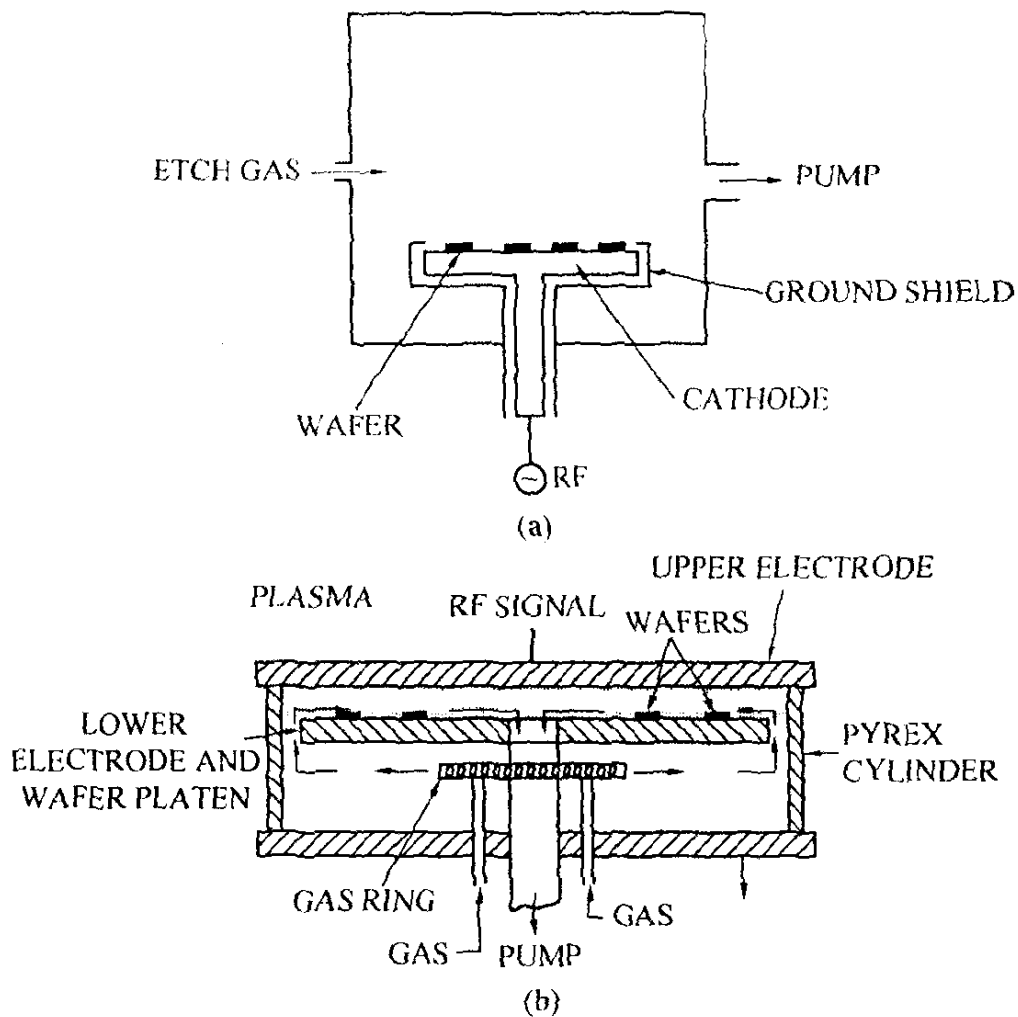


Figure 3 - 5 Dry-etching system

3.9 Integrated Devices

Microwave, photonic, and power applications generally employ discrete devices. For example, an optical IMPATT diode is used as a microwave generator, an injection laser as an optical source, and a thyristor as a high-power switch. However, most electronic systems are built on the integrated circuit (IC), which is an ensemble of both active (e. g., transistor) and passive devices (e. g., resistor and capacitor) formed on and within a single-crystal semiconductor substrate and interconnected by a

metallization pattern in this chapter we combine the basic processes described in previous chapters to fabricate active and passive components in an IC. Because the key element of an IC is the transistor, specific processing sequences are developed to optimize its performance. We shall consider three major IC technologies associated with the three transistor families: the bipolar transistor, the MOSFET and the MESFET.

Figure 3 – 6 illustrates the interrelationship between the major process steps used for IC fabrication. Polished wafers with a specific resistivity and orientation are used as the starting material. The film formation steps include epitaxially grown semiconductor films; thermally grown oxide films; and deposited polysilicon, dielectric, and metal films. Film formation is often followed by impurity doping by methods such as diffusion and ion implantation or lithography, Lithography is generally followed by etching which in turn is often followed by another impurity doping or film formation. The final IC is made by sequentially transferring the patterns from each mask. Level by level, onto the surface of the semiconductor wafer.

After processing, each wafer contains hundreds of identical rectangular chips (or dice), typically between 1 and 10 mm on each side, as shown in Figure 3 – 7(a). The chips are separated by sawing or laser cutting: Figure 3 – 7(b) shows a separated chip. Schematic top views of a single MOSFET and a single bipolar transistor are shown in Figure 3 – 7(c) to give some perspective of the relative size of a transistor. Defective chips are usually marked with a dab of black ink. Good chips are selected and packaged to provide an appropriate thermal, electrical, and interconnection environment in electronic applications.

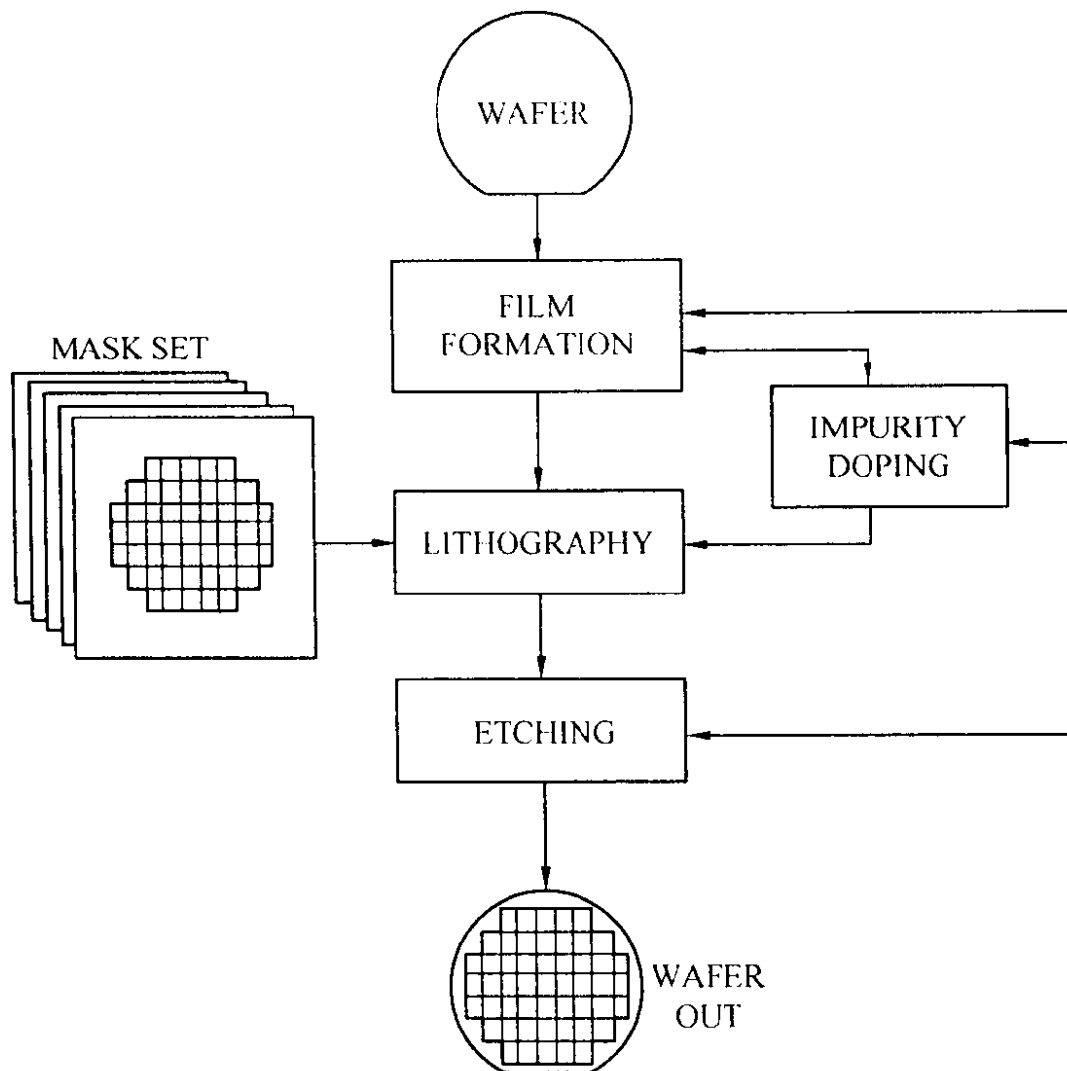


Figure 3 – 6 Schematic flow diagrams of integrated circuit fabrication

Integrated-circuit chips may contain from a few components (transistors, diodes, resistors, capacitors, etc.) to as many as a million or more. Since the invention of the integrated circuit in 1958, the number of components on a state-of-the-art IC chip has grown exponentially. We usually refer to the complexity of an IC as small-scale integration (SSI) for up to 100 components per chip, medium-scale integration (MSI) for up to 1 000 components per chip, large-scale integration (LSI) for up to 100 000 components per chip, and very-large-scale integration (VLSI) for even larger numbers of components per chip. In this chapter we

show two VLSI chips: a 32-bit microprocessor chip, which contains over 150 000 components, and a 1-megabit dynamic-random-access-memory (DRAM) chip, which contains over 2 200 000 components.

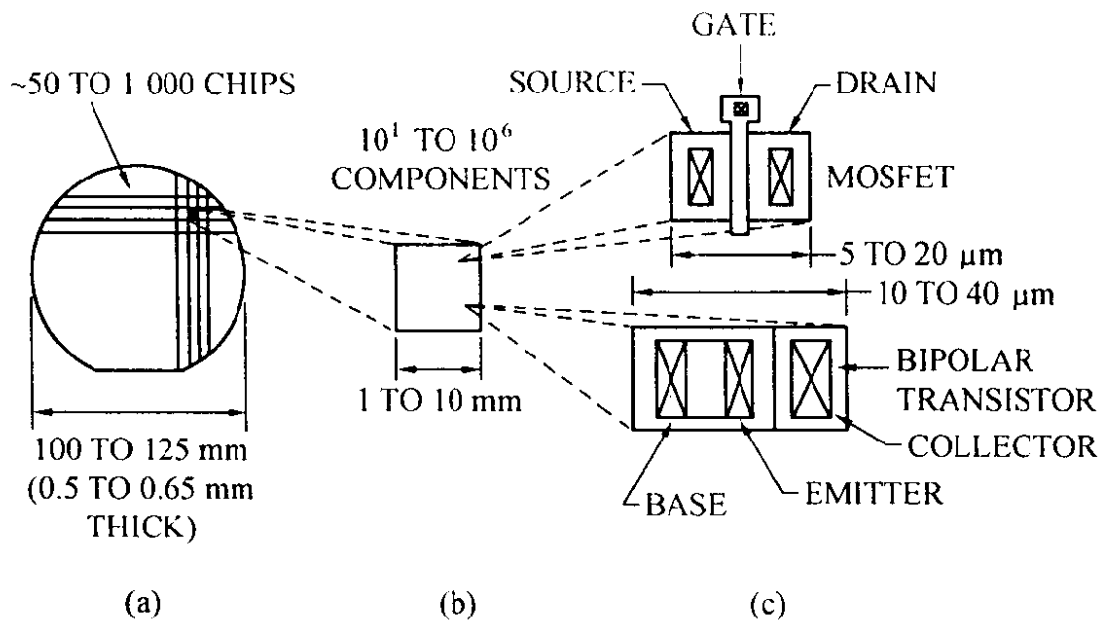


Figure 3 - 7 Size comparison of a wafer to individual components

3.9.1 Passive Components

To form an integrated-circuit resistor, we can define a window in a silicon dioxide layer grown thermally on a silicon substrate and then implant (or diffuse) impurities of the opposite conductivity type into the wafer. Figure 3 - 8 shows the top and cross-sectional views of two resistors; one has a meander shape and the other has a bar shape.

Consider the bar-shaped resistor first. The differential conductance dG of a thin layer of the p-type material that is of thickness dx parallel to the surface and at a depth x (as shown by the B-B cross section) is

$$dG = q\mu_p p(x) \frac{W}{L} dx \quad (3-19)$$

Where W is the width of the bar and L is the length of the bar (we neglect the end contact areas for the time being). The total conductance of the entire implanted region of the bar is given by

$$G = \int_0^{x_j} dG = q \frac{W}{L} \int_0^{x_j} \mu_p p(x) dx \quad (3-20)$$

Where x_j is the junction depth. If the values of μ_p (which is a function of the hole concentration) and the distribution of $p(x)$ are known, the total conductance can be evaluated from Eq. 3-20.

We can write

$$G \equiv g \frac{W}{L} \quad (3-21)$$

Where $g \equiv q \int_0^{x_j} \mu_p p(x) dx$ is the conductance of a square resistor pattern, that is, $G = g$ when $L = W$.

The resistance is therefore given by

$$R \equiv \frac{1}{G} = \frac{L}{W} \left[\frac{1}{g} \right] \quad (3-22)$$

Where $1/g$ usually is denoted by the symbol $R_{\square}$ and is called the sheet resistance. The sheet resistance has units of ohms per square ($\Omega/\square$).

Many resistors in an integrated circuit are fabricated simultaneously by defining different geometric patterns in the mask. Since the same processing cycle is used for all these resistors, it is convenient to separate the resistance into two parts: the sheet resistance $R_{\square}$, determined by the implantation (or diffusion) process; and the ratio L/W , determined by the pattern dimensions. Once the value of $R_{\square}$ is known, the resistance is given by the ratio L/W , or the number of squares (each square

has an area of $W \times W$) in the resistor pattern. The end contact areas will introduce additional resistance to the integrated-circuit resistors. Each end contact corresponds to approximately 0.65 square. For the meander-shaped resistor, the electric-field lines at the bends are not spaced uniformly across the width of the resistor but are crowded toward the inside corner. A square at the bend does not contribute exactly 1 square, but rather 0.65 square. For example, a resistor $90\text{ }\mu\text{m}$ long and $10\text{ }\mu\text{m}$ wide contains 9 squares ($9\text{ }\square$). The two end contacts correspond to $1.3\text{ }\square$. If the implanted layer has a sheet resistance of $1\text{ k}\Omega/\square$, the value of the resistor is $10.3\text{ }\square \times 1\text{ k}\Omega/\square = 10.3\text{ k}\Omega$

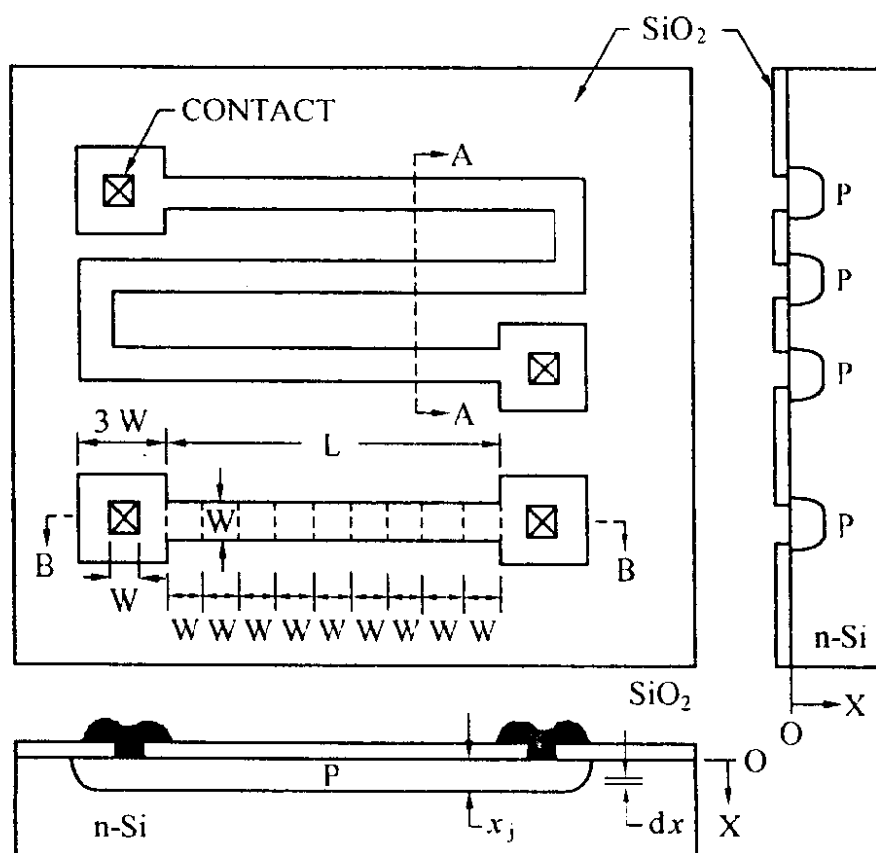


Figure 3 - 8 Integrated circuit resistors

There are basically two types of capacitors used in integrated circuits: MOS capacitors and p-n junctions. The MOS (metal-

oxide-semiconductor) capacitor can be fabricated by using a heavily doped region (such as an emitter region) as one plate, the top metal electrode as the other plate, and the intervening oxide layer as the dielectric. To form an MOS capacitor, a thick oxide layer is thermally grown on a silicon substrate. Next, a window is lithographically defined and then etched in the oxide. Diffusion or ion implantation is used to form a p^+ - region in the window area, while the surrounding thick oxide serves as a mask. A thin oxide layer is then thermally grown in the window area, followed by a metallization step. The capacitance per unit area is given by

$$C = \frac{\epsilon_{ox}}{d} \text{ F/cm}^2 \quad (3 - 23)$$

Where ϵ_{ox} is the dielectric permittivity of silicon dioxide (the dielectric constant ϵ_{ox}/ϵ_0 is 3.9) and d is the thin-oxide thickness. To increase the capacitance further, insulators with higher dielectric constants are being studied (e. g. , Si_3N_4 and Ta_2O_5 , with dielectric constants of 8 and 22, respectively). The MOS capacitance is essentially independent of the applied voltage, because the lower plate of the capacitor is made of heavily doped material. This also reduces the series resistance associated with it.

A p-n junction is sometimes used as a capacitor in an integrated circuit. The detailed fabrication process will be considered in Section 3.8.2, because this structure forms part of a bipolar transistor. As a capacitor, the device is usually reverse-biased, that is, the p-region is reverse-biased with respect to the n^+ -region. The capacitance is not a constant but varies as $(V_R + V_{bi})^{-1/2}$, where V_R is the applied voltage and V_{bi} is the built-in potential. The series resistance is considerably higher than that of MOS capacitor because the p-region has higher resistivity than

does the p^+ -region.

3.9.2 Bipolar Technology

For IC applications, especially for VLSI, bipolar transistors must be reduced in size to meet the high-density requirement. The main differences in a bipolar transistor in an IC compared to a discrete transistor are that all electrode contacts are located on the top surface of the IC wafer, and each transistor must be electrically isolated to prevent interactions between devices. Prior to 1970, both the lateral and vertical isolations were provided by p-n junctions, and the lateral p-isolation region was always reverse-biased with respect to the n-type collector. In 1971, thermal oxide was used for lateral isolation, resulting in a substantial reduction in device size, because the base and collector contacts about the isolation region. In the mid-1970s, the emitter was extended to the walls of the oxide, resulting in an additional reduction in area. At the present time, all the lateral and vertical dimensions have been scaled down and emitter stripe widths have dimensions in the micron region.

The majority of bipolar transistors used in ICs are of the n-p-n type, because the higher mobility of minority carriers (electrons) in the base region results in higher speed performance than can be obtained with p-n-p types. The lateral oxide isolation approach not only reduces the device size but also reduces parasitic capacitance because of the smaller dielectric constant of silicon dioxide. We shall consider the major process steps that are used to fabricate the device.

For an n-p-n bipolar transistor, the starting material is a p-type lightly doped ($\sim 10^{15} \text{ cm}^{-3}$), $\langle 111 \rangle$ -or $\langle 100 \rangle$ -oriented,

polished silicon wafer. Because the junctions are formed inside the semiconductor, the choice of crystal orientation is not as critical as for MOS devices. The first step is to form a buried layer. The main purpose of this layer is to minimize the series resistance of the collector. A thick oxide (0.5 to 1 μm) is thermally grown on the wafer, and a window is then opened in the oxide. A precisely controlled amount of low-energy arsenic ions (~ 30 keV, $\sim 10^{15}$ cm^{-2}) is implanted into the window region to serve as a predeposit (Figure 3 – 9a). Next, a high-temperature (~ 1100 $^{\circ}\text{C}$) drive-in step forms the n^+ buried layer, which has a typical sheet resistance of $20\ \Omega/\square$.

The second step is to deposit an n-type epitaxial layer. The oxide is removed and the wafer is placed in an epitaxial reactor for epitaxial growth. The thickness and the doping concentration of the epitaxial layer are determined by the ultimate use of the device. Analog circuits (with their higher voltages for amplification) requires thicker layers ($\sim 10\ \mu\text{m}$) and lower dopings ($\sim 5 \times 10^{15}$ cm^{-3}), while digital circuits (with their lower voltages for switching) require thinner layers ($\sim 3\ \mu\text{m}$) and higher dopings ($\sim 2 \times 10^{16}$ cm^{-3}). Figure 3 – 9b shows a cross-sectional view of the device after the epitaxial process. Note that there is some outdiffusion from the buried layer into the epitaxial layer. To minimize the outdiffusion, a low-temperature epitaxial process should be employed, and low-diffusivity impurities should be used in the buried layer (e.g., As).

The third step is to form the lateral oxide isolation region. A thin-oxide pad (~ 50 nm) is thermally grown on the epitaxial layer, followed by a silicon-nitride deposition (~ 100 nm). If nitride is deposited directly onto the silicon without the thin oxide

pad, the nitride may cause damages to the silicon surface during the subsequent high-temperature steps. Next, the nitride-oxide layers and about half of the epitaxial layer are etched using a photoresist as mask [Figure 3 – 9(c) and 3 – 9(d)]. Boron ions are then implanted into the exposed silicon areas [Figure 3 – 9(d)].

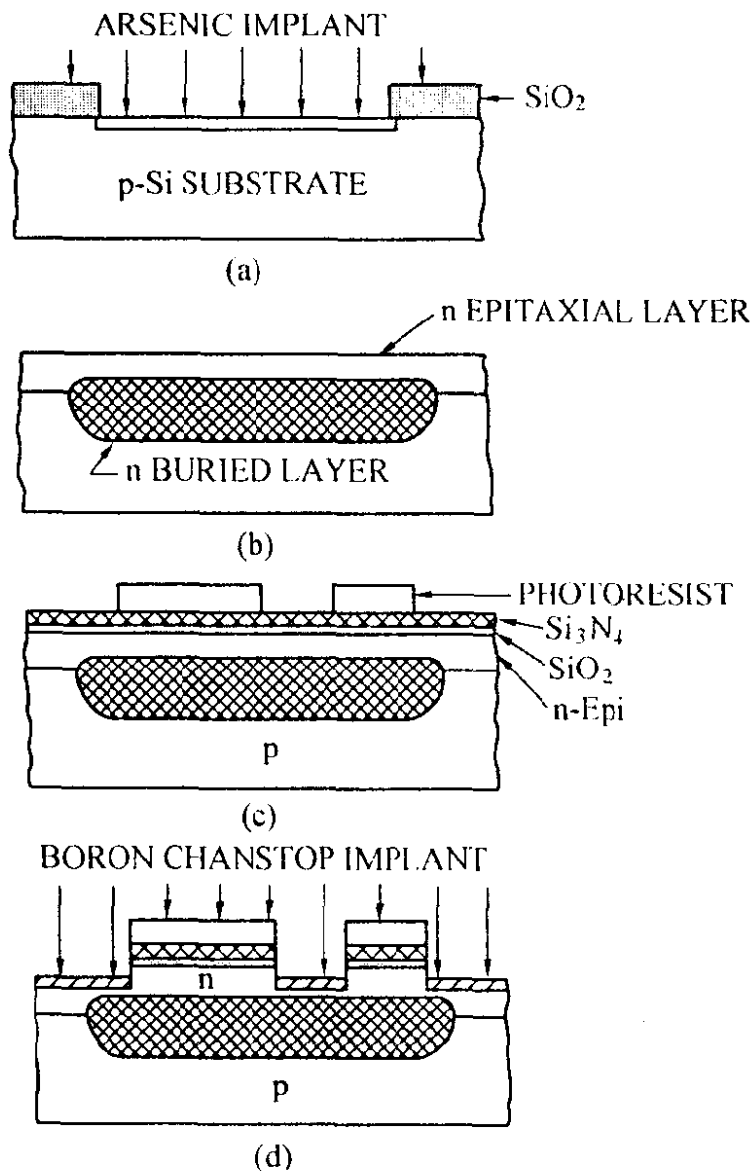


Figure 3 – 9 Cross-sectional views of bipolar transistors fabrication

The photoresist is removed and the wafer is placed in an oxidation furnace. Since the nitride layer has a very low oxidation

rate, thick oxides will be grown only in the areas not protected by the nitride layer. The isolation oxide is usually grown to a thickness such that the top of the oxide becomes coplanar with the original silicon surface to minimize the surface topography. Figure 3-9(a) shows the cross section of the isolation oxide after the removal of the nitride layer. Due to segregation effects, most of the implanted boron ions are pushed underneath the isolation oxide to form a p^+ -layer. This is called p^+ channel stop (or *chanstop*), because the high concentration of p-type semiconductor will prevent surface inversion and eliminate possible high-conductivity paths among neighboring buried layers.

The fourth step is to form the base region. A photoresist is used as a mask to protect the right half of the device; then, boron ions ($\sim 10^{12} \text{ cm}^{-2}$) are implanted to form the base region. Another lithographic process removes all the thin pad oxide except a small area near the center of the base region.

The fifth step is to form the emitter region. The base contact area is protected by a photoresist mask; then a low-energy high-arsenic-dose ($\sim 10^{16} \text{ cm}^{-2}$) implantation forms the n^+ -emitter and the n^+ -collector contact regions. The photoresist is removed; and a final metallization step forms the contacts to the base, emitter, and collector.

In this above basic bipolar process, there are six film formation operations, six lithographic operations, four ion implantation, and four etching operations. Each operation must be precisely controlled and monitored. Failure of any one of the operations generally will render the wafer useless.

The doping profiles of the completed transistor along a coordinate perpendicular to the surface and passing through the

emitter, base and collector are shown in a figure. The emitter profile is quite abrupt due to the concentration-dependent diffusivity of arsenic. The base doping profile beneath the emitter does not decrease smoothly as described by a Gaussian distribution for limited-source diffusion because of the emitter push effect discussed in Section 3.4. The collector doping is given by the epitaxial doping level ($\sim 2 \times 10^{16} \text{ cm}^{-3}$ for a representative switching transistor); however, at larger depths the collector doping concentration increases due to outdiffusion from the buried layer.

In the isolation scheme described for the bipolar transistor, the device is isolated from other devices by the oxide layer around its periphery and is isolated from its common substrate by an $n^+ - p$ junction (buried layer). In high voltage applications, a different approach, called dielectric isolation, is used to form insulating tubs to isolate a number of pockets of single-crystal semiconductors. In this approach the device is isolated from both its common substrate and its surrounding neighbors by a dielectric layer.

The process sequence for the dielectric isolation is shown in a figure. An oxide is thermally grown on a $\langle 100 \rangle$ -oriented silicon wafer, and windows are lithographically defined in the oxide as shown in a figure. An orientation-dependent etching is used to delineate the V-shaped grooves using the oxide as an etching mask. After removal of the mask, a thermal oxide is grown across the wafer to a thickness of typically $1 \mu\text{m}$. This oxide layer serves as the dielectric isolation between the single-crystal and the polycrystalline silicon, which is subsequently deposited to a thickness of about $250 \mu\text{m}$. The polysilicon can be deposited by

thermal reduction of SiHCl_3 in hydrogen or by thermal decomposition of silane as described in Section 3.2. Finally, the single-crystal side of the wafer is lapped down to give a structure. The resulting wafer consists of the desired tubs of single-crystal silicon isolated from each other by a layer of dielectric (silicon dioxide). Various types of devices can be fabricated within these tubs.

The main advantage of this technique is its higher degree of isolation compare to the lateral oxide isolation technique. The main disadvantages of dielectric isolation are the requirement of precise mechanical alignment of the wafer during the lapping operation and the warping of the wafer caused by the high-temperature deposition of polysilicon. Because of these disadvantages and the added process steps requited to form the isolated tubs, the dielectric isolation approach is mainly used for high-voltage ICs or for circuits which must be insensitive to radiation. This approach is valuable for such circuits since all the electron-hole pairs generated beneath the insulator by high-energy radiations cannot participate in the operation of the devices located in the isolated tubs.

The processing steps for passive and active devices described in previous sections can now be combined to show how various integrated circuits are fabricated. As an example, we shall consider the bipolar-inverter circuit, which is one of the basic elements of most digital systems. An inverter causes its output to be in the logic state opposite that of its input. In other words, when the input is at logic 0, the output is at logic 1 or vice versa. In terms of voltage levels and positive logic, this means that with a low-voltage level at the input, the output is at a high-voltage level,

or vice versa. A representative circuit of a bipolar-transistor logic inverter is shown in Figure 3 – 10a. A perspective view of an IC realization of the circuit is shown in Figure 3 – 10b. The bipolar transistor and the two resistors are fabricated using the lateral oxide isolation technique. For clarity, the intermediate dielectric layers are not shown.

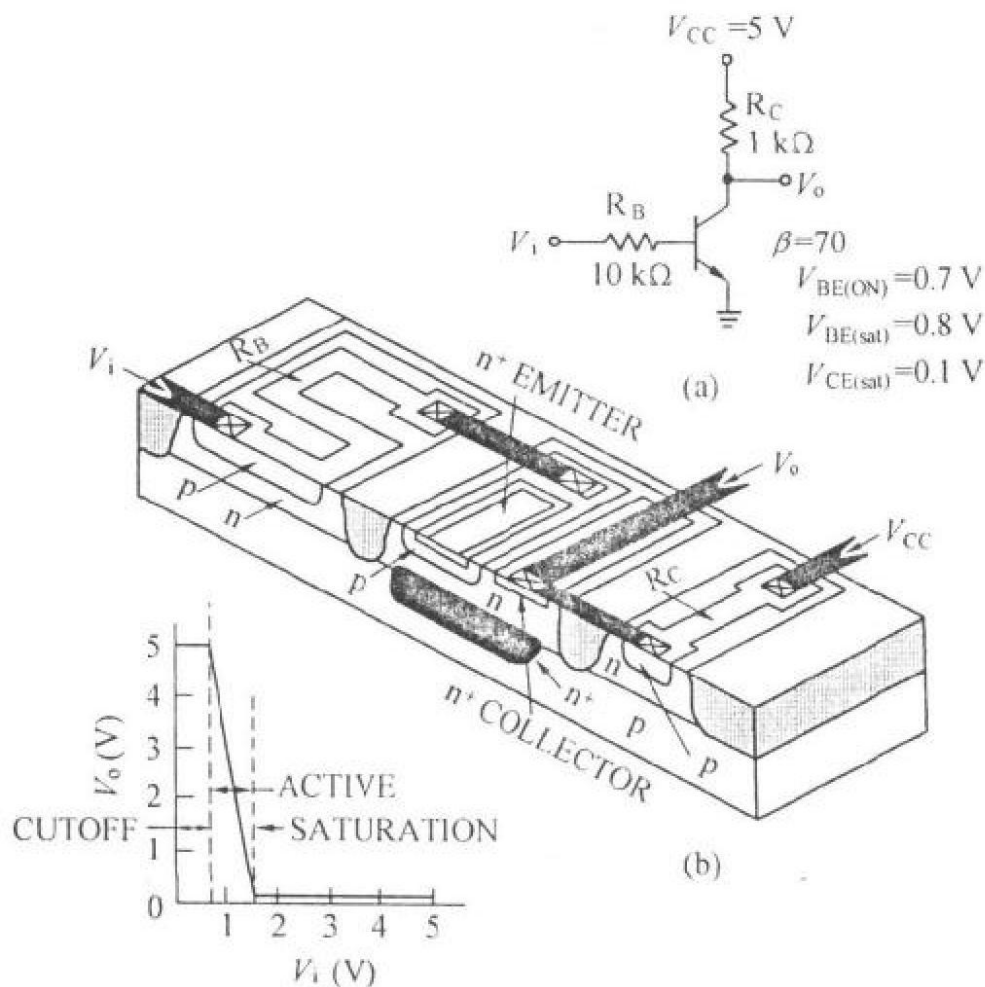


Figure 3 – 10 Bipolar transistor inverter

Figure 3 – 10c shows the voltage transfer characteristic of the bipolar inverter. The transfer characteristic relates the output voltage to the input voltage. When the input voltage (V_i) is less than the turn-on voltage, $V_{BE(on)}$, the transistor is in the cutoff region; and the collector current I_C is essentially zero and the

output voltage V_o will be equal to $(V_{CC} - I_C R_C) \approx V_{CC} = 5 \text{ V}$. When V_i is increased to a value above $V_{BE(on)}$, the transistor turns on and enters the forward active region where the collector current is related to the base current as $I_C \approx \beta_0 I_B R_C$. Thus, the direction of the voltage change is inverted. With sufficient input voltage, the output voltage will decrease until the transistor enters into the saturation region. In the saturation region, the output voltage will remain at a constant low value as the input voltage is further increased.

Integrated injection logic (I^2L) is extensively used in IC logic and memory designs. Attractive features of I^2L include compatibility with bipolar transistor processing, ease of layout, and high packing density. The basic logic cell has a lateral p-n-p transistor (Q_1) and an inverted vertical n-p-n transistor (Q_2) with multiple collector contacts.

In the lateral p-n-p transistor, the p-type emitter and the collector regions are formed simultaneously during the base implantation (or diffusion) step of the n-p-n transistor fabrication. The n-epitaxial layer provides the base region. Since the current flow is mainly along the lateral direction, this device is called a lateral transistor. The advantage of the lateral transistor is that it does not require additional processing steps beyond those necessary for the fabrication of the standard n-p-n transistor. However, the performance of the lateral transistor is inferior to that of the vertical n-p-n transistor, because some of the injected carriers from the emitter will flow vertically toward the n^+ -buried layer. These carriers recombine in the base and are not available at the collector; therefore, the current gain is lower.

For the inverted vertical n-p-n transistor, the buried layer

serves as its emitter, the collection region of the lateral p-n-p transistor serves as its base, and the top n^+ -regions serve as its multiple collectors. With node E at a fixed positive bias, and with a high input voltage (logic 1) applied to V_i , additional current is injected from the emitter of the lateral p-n-p transistor (Q_1) into its collector, which is also the base of Q_2 , thus driving Q_2 into a saturation condition. This, in turn, causes the outputs of Q_2 (V_o) to drop to a logic 0. Therefore, the I^2L can perform the functions of a logic inverter. Since I^2L does not require resistors or isolation regions between the multiple collectors and common emitter, its circuit density can be very high. The I^2L structure uses lateral oxide isolation and its fabrication is compatible with the bipolar processing described in Section 3.8.1.

3.9.3 MOSFET Technology

At present, the MOSFET is the dominant device used in VLSI circuits because it can be scaled to smaller dimensions than can other types of devices. MOSFET technology can be subdivided into NMOS (n-channel MOSFET) technology and CMOS (complementary MOSFET) technology, which provides n-channel and p-channel MOSFETs on the same chip. Both technologies are attractive because the NMOS circuit has fewer processing steps than does the bipolar transistor, while the CMOS circuit has substantially lower power consumption compared to both bipolar and NMOS circuits.

In the early 1970s, the gate length was $7.5 \mu\text{m}$ and the corresponding device area was about $6\,000 \mu\text{m}^2$. As the device is scaled down, there is a drastic reduction in the device area. For a MOSFET with a gate length of $1 \mu\text{m}$, the device area shrinks to

less than 1% of the early MOSFET. We expect that device miniaturization will continue. We consider the fundamental limits of the devices in section 3.9.

Figure 3 - 11 shows a perspective view of an n-channel MOSFET prior to its final metallization. The top layer is a phosphorus-doped silicon dioxide (P-glass), which is used as an insulator between the polysilicon gate and the gate metallization. Compare Figure 3 - 11 with Figure 3 - 10 for the bipolar transistor, and note that a MOSFET is considerably simpler in its basic structure. Although both devices use lateral oxide isolation, there is no need for vertical isolation in the MOSFET, while a buried layer n^+ -p junction is required in the bipolar transistor. The doping profile in a MOSFET is not as complicated as that in a bipolar transistor, and the control of the dopant distributions is also less critical. We shall consider the major process steps that are used to fabricate the device of Figure 3 - 11.

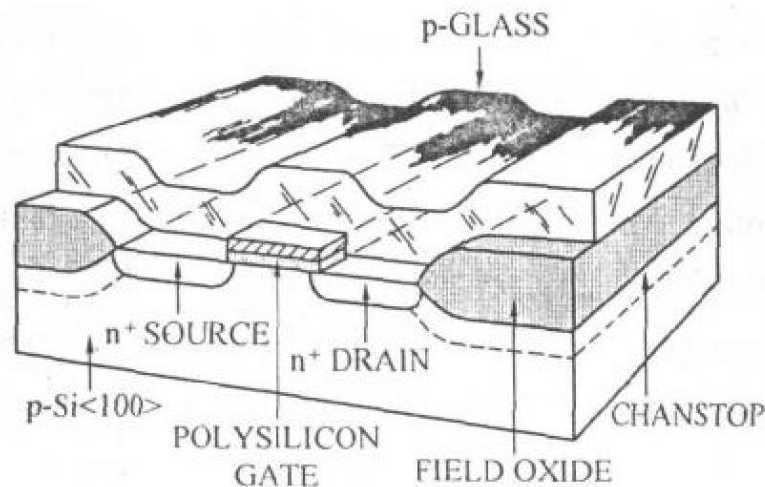


Figure 3 - 11 Perspective view of a n-channel MOSFET

For NMOS processing, the starting material is a p-type lightly doped ($\sim 10^{15} \text{ cm}^{-3}$), $\langle 100 \rangle$ -oriented, polished silicon wafer. The $\langle 100 \rangle$ -orientation is preferred over $\langle 111 \rangle$ because it has

an interface trap density which is about one tenth that of $\langle 111 \rangle$. The first step is to form the oxide isolation region. The process sequence for this step is similar to that for bipolar transistor. A thin oxide pad (~ 50 nm) is thermally grown, followed by a silicon nitride (~ 10 nm) deposition. The active device area is defined by a photoresist mask, and a boron chanstop layer is then implanted through the composite nitride - oxide layer. The nitride layer not covered by the photoresist mask is subsequently removed by etching. After stripping the photoresist, the wafer is placed in an oxidation furnace to grow an oxide (called the field oxide) where the nitride layer is removed, and to drive in the boron implant. The thickness of the field oxide is typically 0.5 to 1 μm .

The second step is to grow the gate oxide and to adjust the threshold voltage. The composite nitride - oxide layer over the active device area is removed, and a thin gate oxide layer (a few hundred angstroms) is grown. For an enhancement-mode n-channel device, boron ions are implanted in the channel region to increase the threshold voltage to a predetermined value (e.g., + 0.5 V). For a depletion-mode n-channel device, arsenic ions are implanted in the channel region to decrease the threshold voltage (e.g., -0.5 V).

The third step is to form the gate. A polysilicon layer is deposited and is heavily doped by diffusion or implantation of phosphorus to a typical sheet resistance of 20 to 30 $\Omega/\square$. This resistance is adequate for MOSFETs with gate lengths larger than 3 μm . For smaller devices, refractory metals (e.g., Mo) or polycides (i.e., a composite layer of metal silicide and polysilicon) can be used as the gate materials, to reduce the sheet resistance to about 1 $\Omega/\square$.

The fourth step is to form the source and drain. After the gate is patterned, it serves as a mask for the arsenic implantation (~ 30 keV, $\sim 10^{16} \text{ cm}^{-2}$) to form the source and drain, which are self-aligned with respect to the gate. At this stage, the only overlapping of the gate is due to lateral straggling of the implanted ions (for 30 keV As, $\Delta R_{\perp}$ is only 5 nm). If low-temperature processes are used for subsequent steps to minimize lateral diffusion, the parasitic gate-drain and gate-source coupling capacitances can be much smaller than the gate-channel capacitance.

The last step is the metallization. A phosphorus-doped oxide (P-glass) is deposited over the entire wafer and is flowed by heating the wafer to give a smooth surface topography. Contact windows are defined and etched in the P-glass. A metal layer, such as aluminum, is then deposited and patterned. A cross-sectional view of the completed MOSFET is shown in Figure 3 – 12 (c), and the corresponding top view is shown in Figure 3 – 12(d). The gate contact is usually made outside the active device area to avoid possible damage to the thin gate oxide.

In this NMOS process, there are six film-formation operations, four lithographic operations, three ion imputations, and four etching operations. There are savings of two lithographic operations and one implantation compared to the basic bipolar process.

The NMOS Logic Gate: Figure 3 – 13(a) shows the circuit of a basic logic gate, the two-input NOR gate, which has two enhancement mode MOSFETs (these devices are called drivers) and one depletion mode MOSFET (called the load). The layout of the NOR gate is shown in Figure 3 – 13(b), and a cross-sectional

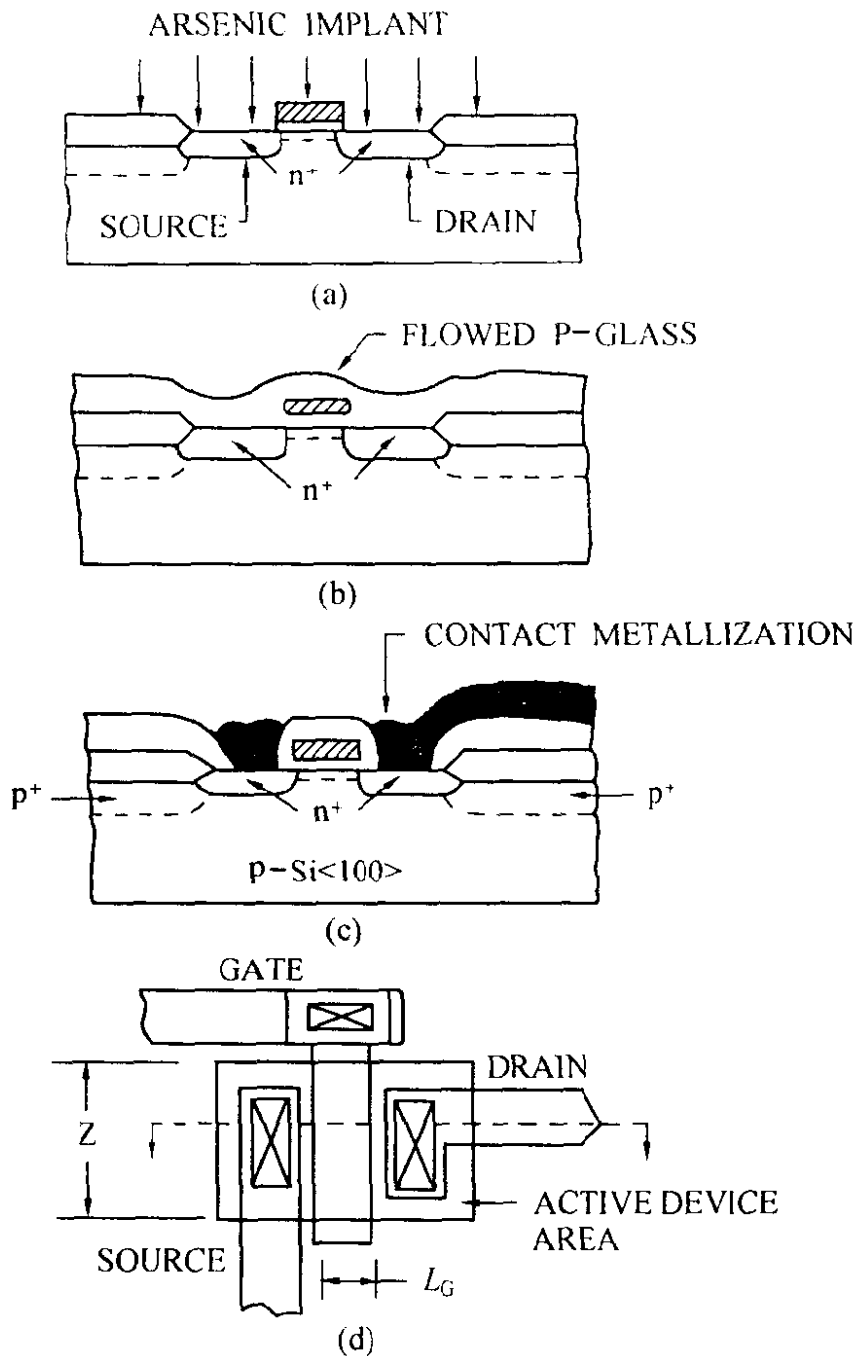


Figure 3 - 12 NMOS fabrication sequence

view of the gate along the line AA' is shown in Figure 3 - 13(c). The cross-sectional view also corresponds to a depletion load inverter. When there is a logic 0 input signal (low voltage at V_i) at the driver device, the device has high impedance (with very small channel current), and the output is at logic 1 (close to V_{DD}). On the other hand, when the input signal is at logic 1, the driver

conducts a large current with a small voltage drop across the device, and the output is at logic 0. The input and output signals are inverted, and the transfer characteristic is similar to that of Figure 3 - 10(c). For the two-input NOR gate [Figure 3 - 13(a)], a logic 1 output signal can be obtained if, both inputs have logic 0 signals. For any other combinations of input signals, the output will have a logic 0 signal.

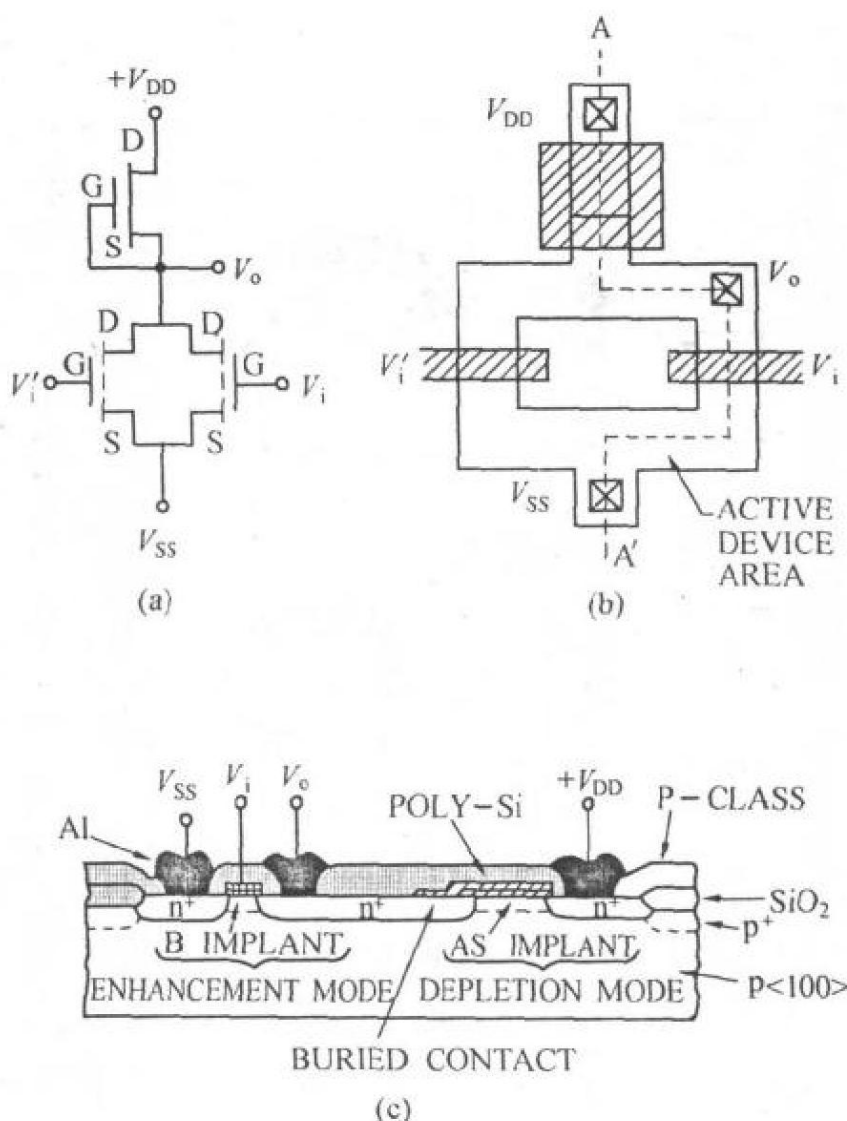


Figure 3 - 13 NMOS low-input NOR gate

The processing steps are essentially the same as described in next section. The enhancement mode MOSFETs (with

corresponding input voltages of V_i) and V'_i) are formed by implanting boron ions in the channel region to control the threshold voltage; while the depletion mode MOSFET is formed by implanting arsenic ions in the channel region to reduce the threshold voltage. The gate electrode of the depletion mode MOSFET is connected to its source by means of a buried contact as shown in Figure 3 – 13(c). Similarly, many other logic circuits can be implemented using NMOS technology.

NMOS Memory Devices: Memories are devices which can store digital information (or data) in terms of bits (binary digits). Various memory chips have been designed and fabricated using NMOS technology. For most large memories, the random access memory (RAM) organization is preferred. In a RAM, memory cells are organized in a matrix structure, and they can be accessed in random order, independent of their physical locations, to store (write) or to retrieve (read) data. A static random-access memory (SRAM) can retain stored data indefinitely. The SRAM can be implemented as a flip-flop circuit to store one bit of information. A SRAM cell has four enhancement mode MOSFETs and two depletion mode MOSFETs. The depletion mode MOSFETs can be replaced by resistors formed in undoped polysilicon to minimize power consumption.

To further reduce area and power consumption, dynamic random-access memory (DRAM) was developed. Figure 3 – 14(a) shows the circuit diagram of the one-transistor DRAM cell in which the transistor serves as a switch, and one bit of information can be stored in the storage capacitor. The voltage level on the capacitor determines the state of the cell. For example, + 5 V may be removed (typically in a few milliseconds) by the leakage currents

of the capacitor; thus dynamic memories require periodic “refreshing” of the stored charge. Because of its small cell area and low power consumption, the DRAM has the highest component density per chip.

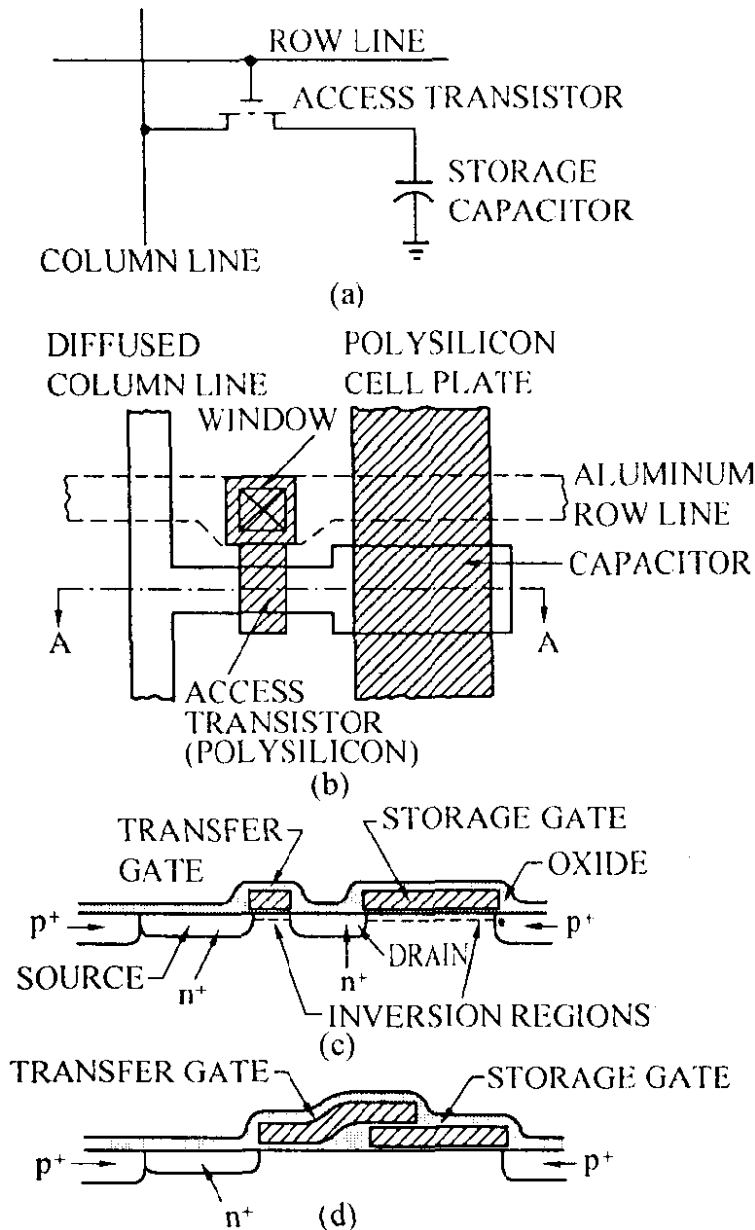


Figure 3 - 14 Single-transistor dynamic-RAM cell with a storage capacitor

Figure 3 - 14 (b) shows the layout of a DRAM cell, and Figure 3 - 14 (c) shows the corresponding cross section through AA. The storage capacitor uses the channel region as one plate, the polysilicon gate as the other plate, and the gate oxide as the

dielectric. The row line is an aluminum track to minimize RC delays. The column line is formed by n^+ -diffusion. The internal drain region of the MOSFET serves as a conductive link between the inversion layers under the storage gate and the transfer gate. This drain region can be eliminated by using the double-level polysilicon approach shown in Figure 3 - 14 (d). The second polysilicon electrode is separated from the first polysilicon capacitor plate by an oxide layer that is thermally grown on the first-level polysilicon before the second electrode has been defined. The charge from the column line can therefore be transmitted directly to the area under the storage gate by the continuity of inversion layers under the transfer and storage gates.

Figure 3 - 15 (a) shows a CMOS inverter. The gate of the upper PMOS device is connected to the gate of the lower NMOS device. Both devices are enhancement mode MOSFETs with threshold voltages less than zero for the PMOS device and V_{Th} greater than zero for the NMOS device (typically $V_{Tp} = -0.5$ to -1.0 V, and $V_{Th} = 0.5$ to 1.0 V). When the input voltage V_i is at ground or at ground or at small positive values, the PMOS device is turned on (the gate-to-ground potential of PMOS is $-V_{DD}$ the PMOS (with $V_{GS} = 0$) is turned off, and the NMOS is turned on ($V_i = V_{DD} > V_{Th}$). Therefore, the output voltage V_o equals zero (logic 0). The behavior of the CMOS inverter has a transfer characteristic similar to that of other inverters described previously. However, the CMOS inverter has a unique feature: in either logic state, one device in the series path from V_{DD} to ground is nonconductive. The current that flows in either steady state is a small leakage current; and only when both devices are on during switching does a significant current flow through the CMOS

inverter. Thus, the average power dissipation is small, in the order of nanowatts. As the number of components per chip increases, the power dissipation becomes a major limiting factor. The low power consumption the most attractive feature of the CMOS circuit.

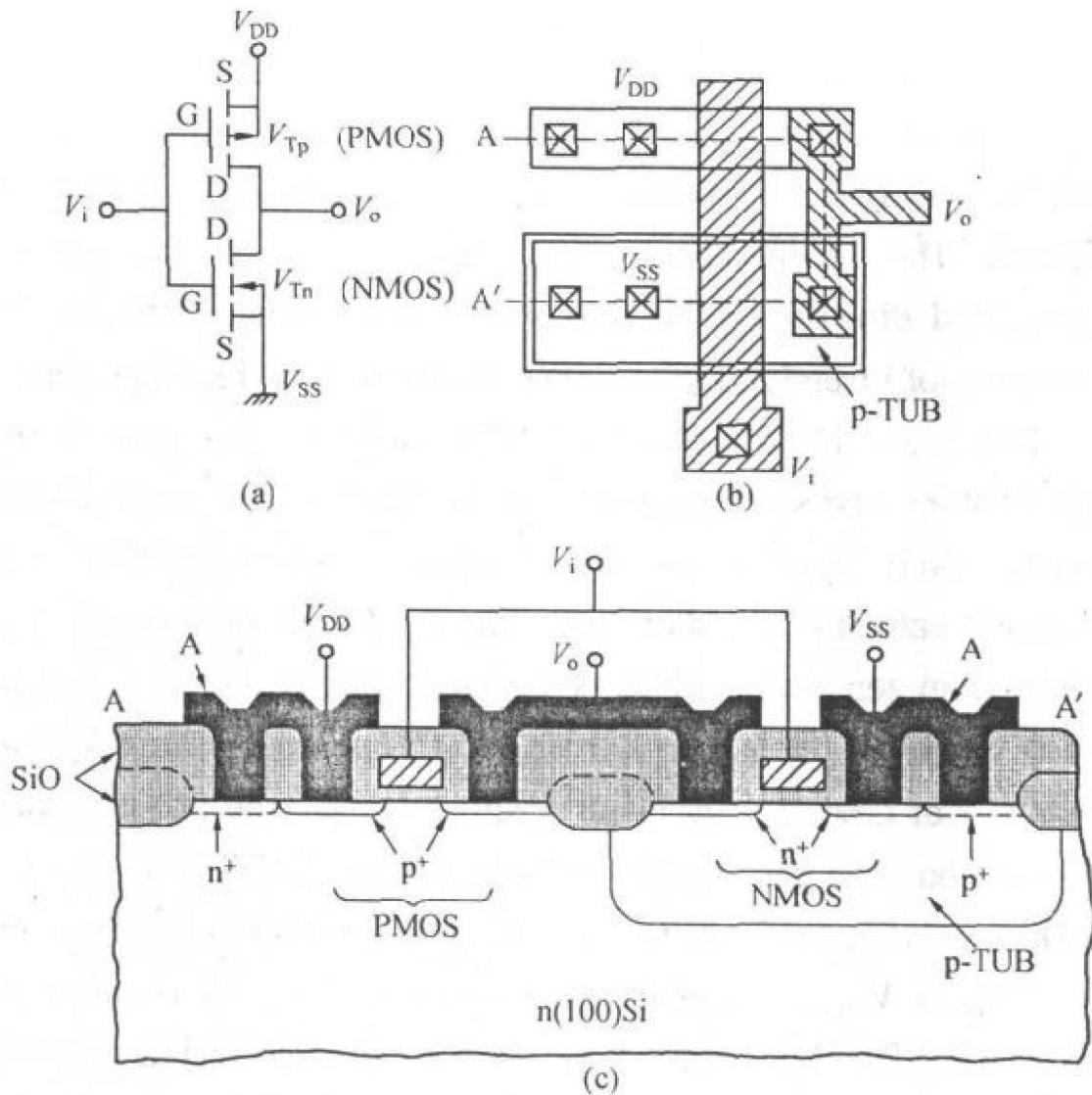


Figure 3 - 15 CMOS inverter

Figure 3 - 15(b) shows a layout of the CMOS inverter, and Figure 3 - 15(c) shows the device cross section along the A—A' line. In the processing, a p-tub (also called a p-well) is first implanted and subsequently driven into the n-substrate. The p-type

dopant concentration must be high enough to over compensate the background doping of the n-substrate. The subsequent processes for the n-channel MOSFET in the p-tub are identical to those described previously. For the p-channel MOSFET, $^{11}\text{B}^+$ or $^{49}(\text{BF}_2)^+$ ions are implanted into the n substrate to form the source and drain regions. A channel implant of $^{75}\text{As}^+$ ions may be used to adjust the threshold voltage, and an n^+ -chanstop is formed underneath the field oxide around the p-channel device. Because of the p-tub and the additional steps needed to make the p-channel MOSFE, the number of steps to make a CMOS circuit is essentially double that to make an NMOS circuit. Thus, we have a trade-off between the complexity of processing and reduction in power consumption.

Instead of the p-tub described above, an alternate approach is to use an n-tub formed in p-type substrate. In this case, the n-type dopant concentration must be high enough to overcompensate for the background doping of the p-substrate (i.e., $N_D > N_A$). In both the p-tub and the n-tub approach, the channel mobility will be degraded, because mobility is determined by the total dopant concentration ($N_A + N_D$). A recent approach uses two separate tubs implanted into a lightly doped substrate. This is called the twin tub approach. Because no overcompensation is needed in either of the twin tubs, higher channel mobilizes can be obtained.

All CMOS circuits have the potential for a troublesome problem called *latchup* that is associated with parasitic bipolar transistors. Note that an n-p-n transistor can be formed with an n^+ -source or drain as its emitter, the p-tub as its base, and the adjacent n-tub as its collector. Similarly, a p-n-p transistor can be formed with a p^+ -source or drain as its emitter and with the n-tub

and p-tub as its base and collector, respectively. These two transistors can be coupled together to act as a thyristor. If the product of the current gains of the two transistors exceeds unity, a large current can flow between V_{DD} and V_{SS} , which is the phenomenon known as latchup. Because of the high current, latchup can cause permanent damage to CMOS circuits.

To avoid latchup, we must reduce the current gain of the parasitic bipolar transistors. One method is to use gold doping or neutron irradiation to lower the minority carrier lifetimes. However, this approach is difficult to control and increases the leakage current. An effective technique is to use the trench isolation. In this technique a trench is formed in the silicon by anisotropy reactive-sputter etching. An oxide layer is thermally grown on the bottom and walls of the trench, which is then refilled by deposited polysilicon or silicon dioxide. This technique can eliminate latchup because the n-channel and p-channel devices are physically isolated by the refilled trench.

A 100-mm silicon wafer containing forty eight 32-bit microprocessor chips and eight test chips is shown. From the identifying flats on the wafer, we recognize that it is a $\langle 100 \rangle$ -oriented n-type wafer. The devices are fabricated using twin-tub 2- μm CMOS design rules. The area of each chip is about 100 mm^2 , containing 150 000 components. The chip can perform 1 million instructions per second, and it consumes 700mW power at full speed.

A pair of these devices forms a static memory cell that resembles the static RAM with the polysilicon resistor load cell. Such a static RAM substantially reduces the cell area while it offers even lower standby power.

Another potentially useful CMOS technology is the silicon-on-insulator (SOI) structure. A cross section of CMOS on SOI is illustrated in Figure 26b. One technique to form such a structure is to grow a layer of silicon dioxide on silicon to a thickness of about $1\text{ }\mu\text{m}$. A polysilicon layer is then deposited on the silicon dioxide and is subsequently recrystallized with laser or incoherent light sources. Semiconductor islands are patterned and defined in the recrystallized silicon. MOSFETs are then fabricated as previously described.

The SOI approach is attractive because it offers high component density and complete immunity to latchup. Circuit performance of SOI is also expected to be superior to conventional CMOS due to reduced parasitic capacitance. The mobility degradation in recrystallized materials has a minimal effect on device performance, since the device speed is determined by the saturation velocity, which is insensitive to defects or crystal imperfection.

3.10 Fundamental Limits of Integrated Devices

Since the beginning of the integrated-circuit year in 1958, the minimum device dimension has been reduced at an annual rate of about 13%. At this rate, the minimum device dimension will shrink to $0.5\text{ }\mu\text{m}$ in the year 1990 and to $0.1\text{ }\mu\text{m}$ around 2000. Assuming that the advanced technology required to fabricate submicron or even smaller devices can be developed, what are the fundamental limits of device dimensions based on physical laws? We shall consider these fundamental limits in terms of intrinsic device limitations, wiring limitations, and power limitations. We shall be mainly concerned with MOSFET-type devices, since

MOSFET is the dominant technology for VLSI applications.

3.10.1 Intrinsic Device Limitations

We first consider the quantum limit, which states that a physical operation performed in a time τ must involve an energy

$$E \geq \frac{\hbar}{\tau} \quad (3 - 24)$$

Where $\hbar$ is the reduced Planck constant. This energy is dissipated as heat. The power dissipated during the operation is

$$P = \frac{E}{\tau} \geq \frac{\hbar}{\tau^2} \quad (3 - 25)$$

Which is a lower limit for power dissipation per unit operation. For an operation speed of 10 ps (10^{-11}) s, the minimum energy dissipated in a switching device as obtained from Eq. 3 - 24 is in the order of 10^{-23} J per operation. The actual value for a MOSFET is about 10^{-14} J, which is quite remote from this quantum limit.

Another intrinsic device limit is the thickness of the gate oxide. When the gate oxide is scaled below 5 nm, there is a finite probability that electrons will pass through the gate oxide by a quantum-mechanical tunneling process. For proper device operation, this tunneling current must be small. Therefore, the tunneling effect sets a fundamental lower limit for the thickness of the gate oxide to about 5 nm.

Semiconductor material properties such as critical electric field and saturation velocity will impose limits on device operation. Consider the minimum propagation time in a cube of silicon material of dimension Δx . If the electrons are traveling at their saturation velocity v_s , the time τ needed to traverse Δx is given by

$$\tau = \frac{\Delta x}{v_s} \quad (3-26)$$

If we assume that the minimum voltage corresponds to the thermal energy divided by the elementary charge (i. e. , $\Delta V = kT/q$) and we use $\epsilon_c = 5 \times 10^5$ V/cm as the critical field and $v_s = 10^7$ cm/s as the saturation velocity in silicon.

Therefore the critical field and saturation velocity limit the minimum transit time in silicon to 5×10^{-15} s. Since gallium arsenide has comparable values of ϵ_c and v_s , the minimum transit time due to material limitations is also about 5×10^{-15} s.

Because of thermal fluctuations, a semiconductor device may randomly switch from a logic 1 to a logic 0, or vice versa. Therefore, the thermal limit of the switching energy is kT , and the operating voltage of a device should be many times kT/q .

The propagation delay (transit time) versus the switching power required per bit of information due to limits imposed by the quantum ($\hbar/\tau$), thermal (kT), and $0.5 \mu\text{m}$ (circles). And material (5×10^{-15} s) limitations. Also shown are the performances of ring oscillators having minimum feature lengths of 2 to $5 \mu\text{m}$, $1 \mu\text{m}$ (dots), and $0.5 \mu\text{m}$ (circles). The power-delay product (this corresponds to the switching energy) for $0.5\text{-}\mu\text{m}$ devices is about 5×10^{-14} J. By scaling the device dimensions and using advanced CMOS technology to minimize power, we may reach $0.1 \mu\text{m}$ with a switching energy three orders of magnitude lower (to 5×10^{-17} J), smaller devices might be fabricated, but at the cost of lowering the voltage or the temperature at which the devices have to be operated.

3.10.2 Wiring Limitations

For intrachip connections (interconnections of components

within a chip) and interchip connections (interconnections between chips), there are three limitations when we shrink the interconnect wirings: electromigration limit, edge capacitance, and wire resistance.

The electromigration problem becomes worse as the minimum device dimension gets smaller. The current from a MOSFET varies as $(1/L)^n$, where L is the channel length and n is between zero and 1 depending on the carrier velocity variation with electric field. If the thickness of a conductive runner is maintained, its area varies as L , so the current density varies as $(1/L)^{n+1}$. A typical 1- μm MOSFET gate can output 1mA, so that a 1- μm square line will carry 10^4 nm/cm^2 , which is close to the electromigration limit. This can of course be circumvented by making the conductors wider and reducing the circuit packing density.

The capacitance of an interconnect consists of parallel-plate and edge capacitances. The parallel-plate part varies directly with line width, but the edge compote stays constant, causing a saturation in line capacitance as the line width shrinks. At 1- μm line widths, these two components are approximately equal; and for practical field oxide sicknesses (~ 0.3 to 1 μm), the edge capacitance starts to be significant at about 2 μm . A plot of capacitance versus the ration of line-width to oxide thickness is shown in a figure. The dot-dash line shows the parallel-plate component. The solid lines are the total capacitance low. A related problem is that line-to-line capacitance will dominate line-to-substrate capacitances as dimensions shrink, causing a crosswalk problem. This can be minimized by routing ground wires next to signal wires at the expense of an increase in the area.

Since the resistivity of a conductor remains constant, the resistance of wires increases as dimensions shrink. This increased resistance coupled with the previously mentioned capacitance result in the increase of the RC time constant, which will limit device speed. All of the wiring-related limitations can be viewed as design constraints, but they are fundamental ones and they detract from the advantages of reducing device dimensions.

3.10.3 Power Limitations

The power required merely to charge and discharge circuit nodes in an integrated circuit is proportional to the number of gates and the frequency at which they are switched (clock frequency). The power can be expressed as $P \cong 1/2 CV^2 nf$, where C is the capacitance per device, V is the applied voltage, n is the number of devices per chip, and f is the clock frequency. The temperature rise caused by this power dissipation in an IC package is limited by the thermal conductivity of the package material, unless auxiliary liquid or gas cooling is used. The maximum allowable temperature rise is limited by the bandgap of the package material, unless auxiliary liquid by the bandgap of the semiconductor ($\sim 100^\circ\text{C}$ for Si with a bandgap of 1.1 eV). For such a temperature rise, the maximum power dissipation of a typical high-performance package is about 10W. As a result, we must limit either the maximum clock rate or the number of gates on a chip. As an example, in an IC containing $1\text{-}\mu\text{m}$ NMOS devices with $C = 5 \times 10^{-3}$ pF, running at a 2-GHz clock rate, the maximum number of gates we can have is about 10^5 if we assume a 10% duty cycle. Again, this can be looked at as a design constraint fixed by basic material parameters.

Ultimate Device Limits. If new device structures are allowed, several basic reasons may limit the smallest dimension to about 100 Å. One is that this dimension corresponds to the spacing between dopant atoms at the degeneracy limit ($\sim 10^{19} \text{ cm}^{-3}$ for Si). Another is that our ability to define a line by any imaginable lithographic means is limited by basic particle scattering considerations to about 10 nm. Certainly, statistical considerations of device geometry and ion dose variation will prohibit any reduction beyond this point.

As we discussed previously, the minimum channel length for MOSFETs is probably around $0.1 \mu\text{m}$ (100 nm). Similar limits can be placed on the minimum feature size in bipolar transistors and MESFETs. Because the fabrication tools will soon be available, we expect that miniaturization will reach these limits in the next two decades.

Beam fabrication techniques such as electron beam and ion beam resistless processes will eventually allow devices with minimum feature lengths approaching 10 nm. However, even to enter the regions below 100 nm will require that we use quantum theory rather than Boltzmann statistics and that we devise radically new device structures (such as molecular electronic structures).

NOTES

In addition, the puller has an overall microprocessor-based control system to control process parameters such as temperature, crystal diameter, pull rate, and rotation speeds, as well as to permit programmed process steps.

另外,有一套基于微处理器控制的控制系统对提拉装置进行全面控制,可控制诸如温度、晶体直径、牵拉速度、旋转速率等,同

时可对工艺过程进行编程控制。

Dielectric layers such as the deposited 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 gate electrode material in MOS devices, as a conductive material for devices with shallow junctions.

淀积的二氧化硅、氮化硅等电解质层作为导电层之间的绝缘层,可作为扩散和离子注入时的掩蔽层,也可作为掺杂薄膜之上覆盖层以防止掺杂损失,同时可作钝化层防止器件受外界杂质、潮湿空气、划痕等的影响。Polycrystalline silicon(多晶硅)通常写为 polysilicon,它对于浅结器件来说是导电材料,在 MOS 型器件中常用做栅极材料,

Film formation is often followed by impurity doping by methods such as diffusion and ion implantation or lithography, Lithography is generally followed by etching which in turn is often followed by another impurity doping or film formation. The final IC is made by sequentially transferring the patterns from each mask.

薄膜形成之后,常常要利用扩散、离子注入进行杂质的注入,或者进行光刻。光刻之后常常要进行腐蚀然后进行另一种杂质注入或薄膜生成,最终的 IC 是将每一层掩膜版的图形按顺序转移到硅片上形成的。

New Words and Expression

epitaxy *n.* 外延

quartzite *n.* 石英岩,硅岩

chloride *n.* 氯化物
nucleation *n.* 晶核形成
polycrystalline *adj.* 多晶的
crucible *n.* 坩锅
graphite susceptor 石墨基座
encapsulation *n.* 封装
arsine *n.* 三氢砷化, 砷
diluent *n.* 稀释液
dielectric *n.* 电介质
passivation *n.* 钝化
silicide *n.* 硅化物
lithographic *n.* 光刻
plasma *n.* 等离子体
pyrolyze *vt.* 使裂解
inert *adj.* 惰性的
ampule *n.* 一次用量的针剂
diffusion *n.* 扩散
electron beam - annealed 电子束退火
thermionic - emission 热电子发射
proximity effect 邻近效应
diffraction *n.* 衍射
autocatalytic *adj.* 自身催化的
acetic acid 乙酸; 醋酸
isopropyl alcohol 异丙基醇
trapezoidal *adj.* 梯形的
sputter *v.* 溅射
anisotropy *n.* 各向异性
halogen *n.* 卤素
refractory *adj.* 难熔的, 难控制的

latchup *n.* 栅锁效应
degeneracy *n.* 简并性[度]
deposition *n.* 淀积
mask *n.* 掩模版
photoresist *n.* 光刻胶
clean room *n.* 超净间
chip *n.* 芯片
wafer *n.* 硅圆片
thermal oxidation 热氧化
SOI 绝缘层上硅结构
CVD 化学汽相淀积

参 考 文 献

- 1 S M Sze. Semiconductor Devices. Bell Telephone Inc., 1985

Chapter 4

Integrated Circuits

4.1 Introduction

It was suggested in the introduction of this chapter that design automation plays an instrumental role in coping with the dramatic increase in complexity of contemporary integrated circuits. Designing a multimillion transistor circuit and ensuring that it operates correctly when the first silicon returns is a daunting task that is virtually impossible without the help of computer aids and well-established design methodologies. In general, the wide range of tools available to a designer can be subdivided into a number of global classes.

- Analysis and verification tools examine the behavior of a circuit and help determine if the response is within specification bounds.

- Implementation and synthesis methodologies aid the designer in generating and optimizing the circuit schematics or layout.

- Testability techniques provide a combination of design approaches and CAD tools to validate the functionality of a fabricated design.

This chapter presents a brief overview of the leading

approaches in each of those categories. Due to the extensive nature of the field, we do not intend to be comprehensive—doing so would require a textbook of its own. Instead we present a user perspective that provides a basic perception and insight into what is offered and can be expected from the different design methodologies. This information should be sufficient for a novice designer to get started. Readers interested in an in-depth discussion of the mechanics of the various tools and environments are referred to the textbooks published on this topic, a number of which are listed in the references.

4.2 Design Analysis and Simulation

The primary expectation of a designer with respect to design automation is the availability of accurate and fast analysis tools. The first computer-aided design (CAD) tool to gain wide acceptance was the SPICE circuit simulator, which is undoubtedly the most utilized computer aid for the design of digital circuits at present [Nagel75]. Unfortunately, circuit simulation takes into account all the peculiarities and second-order effects of the semiconductor devices and tends to be time-consuming. It is rapidly becoming unwieldy when designing complex circuits, unless one is willing to spend days of computer time. The designer can address the complexity issue by giving up modeling accuracy and resorting to higher representation levels. A discussion of the different abstraction levels available to the designer and their impact on simulation accuracy is the topic of this section.

At this point, it is useful to differentiate between simulation and verification. In the simulation approach, the value of a design parameter such as noise margin, propagation delay, or dissipated

energy is determined by applying a set of excitation vectors to the circuit model of choice and extracting parameters from the obtained signal waveforms. While this approach is very flexible, it has the disadvantage that the results depend strongly upon the choice of the excitations. For instance, the delay of a ripple-carry adder varies widely depending upon the input signal. Identification of the worst-case delay requires a careful choice of the excitation vector, so that the complete ripple-carry path is exercised. Failing to do so produces meaningless results. Verification, on the other hand, attempts to extract the system parameters directly from the circuit description. For instance, the critical path of an adder can be recognized from an inspection of the circuit diagram or a model of it. This approach has the advantage that the result is independent of the choice of excitation vectors and is supposedly fool-proof. On the negative side, it requires an inherent understanding of the circuit design style used. For the example of the ripple-carry adder, determination of the propagation delay requires an understanding of the logic operation of the composing circuitry—for example, dynamic or static logic—and a definition of the term propagation delay. As a second example, it is necessary to identify the register elements first before the maximum clock speed of a synchronous circuit can be determined.

The best way of differentiating among the myriad of simulation approaches and abstraction levels is to identify how the data and time variables are represented—as analog, continuous variables, as discrete signals, or as abstract data models.

4.2.1 Representing Digital Data as a Continuous Entity

Circuit Simulation

In the course of this chapter, we have used circuit simulation extensively to illustrate the basic concepts of digital circuits and to validate our manual models. The assignments at the end of each chapter also rely heavily on circuit simulation. Hence, you should be very familiar with the capabilities and peculiarities of simulation and analysis at this level of detail. Some important properties of circuit simulation are worth summarizing.

- When analyzing a digital network using a circuit simulator, the resulting voltage and current signals are represented as continuous waveforms.

- In a transient analysis, time seems to be a continuous variable, and for all practical purposes can be considered as such. In reality, the simulator executed on a digital computer evaluates only a limited number of time points and obtains the intermediate data points by interpolation.

Executing a transient simulation means solving a set of differential equations at each time point. To make matters worse, the accurate modeling of semiconductor devices requires the introduction of nonlinearities, such as the current equations for bipolar and MOS devices or the diode-capacitance model. All capacitors and current sources in this model display a strong nonlinear behavior. The analytical solution of the set of nonlinear differential equations describing such a network is generally based on iteration that is computationally expensive. At each time step, an initial guess is made of the node values based on the values of

the previous time step. This estimate is iteratively refined until some predefined error criterion is met. The tighter the error-bound, the better the accuracy, but the more iterations needed.

The accuracy of circuit simulation is direct function of the quality and complexity of the device models employed. A number of those models for bipolar, MOS, and GaAs devices were introduced earlier in the text. While most circuit simulators use similar solution techniques, the added value is really situated in the robustness, efficiency, and accuracy of these models. For instance, it is possible to come up with very elaborate and accurate models that are completely useless because of poor convergence behavior.

The computational complexity makes circuit simulation impractical for larger circuits. It is, therefore, mostly used for the analysis of the critical parts of a design. Higher-level simulators are employed for the overall analysis.

Timing Simulation

Substantial effort has been invested over time to decrease the computation time at the expense of generality. Consider an MOS digital circuit. Due to the excellent isolation property of the MOS gate, it is often possible to partition the circuit into a number of sections that have limited interaction. A possible approach is to solve each of these partitions individually over a given time period, assuming that the inputs from other sections are known or constant. The resulting waveforms can then be iteratively refined. This relaxation-based approach has the advantage of being computationally more effective than the traditional technique, since expensive matrix inversions are avoided, but is restricted to MOS circuits [White87]. When the circuit contains feedback

paths, the partitions can become large, and simulation performance degrades. The accuracy of these simulators is generally less than that of the SPICE-class tools, but the execution speed can be one or two orders magnitude higher.

4.2.2 Representing Data as a Discrete Entity

When addressing the complexity issue, the only solution is to compromise accuracy. In digital circuits, we are in general not interested in the actual value of the voltage variable, but only the digital value it represents. Therefore, it is possible to envision a simulator where data signals are either in the 0 or 1 range. Signals that do not comply with either of the above are denoted as x , or undefined.

This tertiary representation $\{0, 1, x\}$ is used extensively in simulators at both the device and gate level. By augmenting this set of allowable data values, we can obtain more detailed information while retaining the capability of handling complex designs. Possible extensions are the Z value for a tri-state node in the high-impedance state, and R and F for the rising and falling transients. Some commercially offered simulators go as far as offering more than a dozen possible signal states.

While substantial performance improvement is obtained by discretizing the data representation space, similar benefits can be obtained by considering time as a discrete variable. Consider the voltage waveform of Figure 4 - 1 that represents the signal at the input of an inverter with a switching threshold V_M . It is reasonable to assume that the inverter output change its value one propagation delay after its input crossed V_M . When one is not strictly interested in the exact shape of the signal waveforms, it is

sufficient to evaluate the circuit only at the interesting time points, t_1 and t_2 . Similarly, the interesting points of the output waveform are situated at $t_1 + t_{pHL}$ and $t_2 + t_{pLH}$. A simulator that only evaluates a gate at the time an event happens at one of its inputs is called an event-driven simulator. The evaluation order is determined by putting projected events on a time queue and processing them in a time-ordered fashion. Suppose that the waveform of Figure 4 – 1 acts an input waveform to a gate. An event is scheduled to occur at time t_1 . Upon processing that event, a new event is scheduled for the fan-out nodes at $t_1 + t_{pHL}$ and is put on the time queue. This event-driven approach is evidently more efficient than the time-step-driven approach of the circuit simulators. To take the impact of fan-out into account, the propagation delay of a circuit can be expressed in terms of an intrinsic delay (t_{in}) and a load-dependent factor (t_l), and can differ over edge transitions.

$$t_{pLH} = t_{inLH} + t_{lLH} \times C_L$$

The load C_L can be entered in absolute terms (in pF) or as a function of the number of fan-out gates.

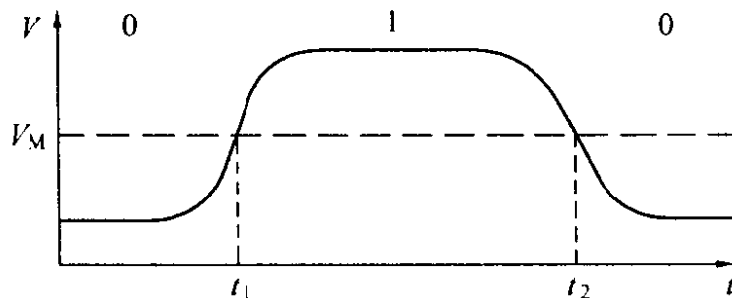


Figure 4 – 1 Discretizing the time variable

While offering a substantial performance benefit, the above approach still has the disadvantage that events can happen at any point in time. Another simplification could be to discretize the

time even further and allow events to happen only at integer multiples of a unit time variable. An example of such an approach is the unit-delay model, where each circuit is attributed a single delay of one unit. Finally, the simplest model is the zero-delay model, in which circuits are assumed to be delay-free. Under this paradigm, time proceeds from one clock event to the next, and all events are assumed to occur instantaneously upon arrival of a clocking event. These concepts can be applied on a number of abstraction levels, resulting in the simulation approaches discussed below.

Switch-Level simulation

The nonlinear nature of semiconductor devices is one the major impediments to higher simulation speeds. The switch-level model [Bryant 81] overcomes this hurdle by approximating the transistor behavior with a linear resistance whose value is a function of the operating conditions. In the off-mode the resistance is set to infinity, while in the on-mode it is set to the average on-resistance of the device (Figure 4 – 2). The resulting network is a time-variant, linear network of resistors and capacitors that can be more efficiently analyzed. Evaluation of the resistor network determines the steady-state values of the signals and typically employs a $\{0, 1, x\}$ model. For instance, if the total resistance between a node and GND is substantially smaller than the resistance to V_{DD} , the node is set to the 0-state. The timing of the events can be resolved by analyzing the RC network. Simpler timing models such as the unit-delay model are also employed.

Gate-level(or Logic) Simulation

Gate-level simulators use the same signal values as the switch-level tools, but the simulation primitives are gates rather

than transistors. This approach enables the simulation of more complex circuits at the expense of detail and generality. For instance, some common VLSI structures such as tri-state busses and pass-transistors are hard to deal with at this level. Since gate level is the preferred design-entry level for many designers, this simulation approach remained extremely popular until the introduction of logic synthesis tools, which moved the focus to the functional or behavioral abstraction layer. The interest in logic simulation was so great that special and expensive hardware accelerators were developed to expedite the simulation process (e.g. [Agrawal90]).

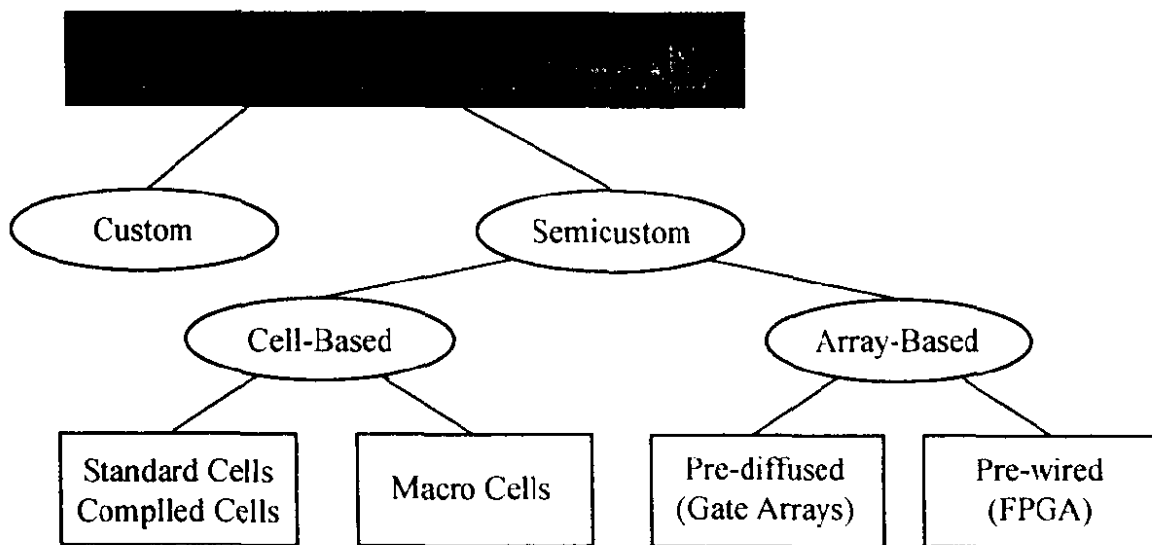


Figure 4 – 2 Switch-level model of CMOS inverter

Functional Simulation

Functional simulation can be considered as a simple extension of logic simulation. The primitive elements of the input description can be of an arbitrary complexity. For instance, a simulation element can be a NAND gate, a multiplier, or an SRAM memory. The functionality of one of these complex units can be described using a modern programming language or a dedicated hardware

description language. For instance, the THOR simulator uses the C programming language to determine the output values of a module as a function of its inputs [Thor 88]. On the other hand, VHDL (VHSIC Hardware Description Language) [VHDL 88] is a specially developed language for the description of hardware designs.

4.2.3 Implementation Approaches

The viability of a microelectronics design depends upon a number of conflicting factors, such as performance in terms of speed of power consumption, cost, and production volume. For example, a microprocessor has to excel in performance at a low cost to the customer to be competitive in the market. Achieving both goals simultaneously is only possible through large sale volumes. The high development cost associated with high-performance design is then amortized over many parts.

Applications such as radar or space systems present another scenario. Since performance is critical, high-performance custom design techniques are desirable. The production volume is small, but the cost of electronic parts is only a small fraction of the overall system and thus not much of an issue. Finally, a majority of the digital designs vie for integration density, not performance. Under these circumstances, the design cost can be reduced substantially by using advanced design-automation techniques that compromise performance but minimize design time. Note that the cost of a semiconductor device is the sum of two components:

- The nonrecurring expense (NRE), which is incurred only once for a design and includes the cost of designing the part
- The production cost per part, which is a function of the process complexity, design area, and process yield.

These economic considerations have spurred the development of a number of distinct implementation approaches that range from high-performance handcrafted design to fully programmable, medium-to-low performance designs.

Custom Circuit Design

When performance of design density is of primary importance, the designer has no other choice than to return to handcrafting the circuit topology and physical design. This approach was the only option in the early days of digital microelectronics. This was adequately demonstrated in the design of the Intel4004 microprocessor, shown on the back cover. The labor-intensive nature of custom design translates into a high cost and a long time-to-market. Therefore, it can only be justified economically under the following conditions:

- The cost can be amortized over a large volume. Microprocessors and semiconductor memories are examples of applications in this cell
- The custom block can be reused many times, for instance as a library cell
- Cost is not the prime design criterion, as it is in space applications

With continuous progress in the design-automation arena, the share of custom design reduces from year to year. Even in high-performance microprocessors, such as the DEC alpha processor, large portions are designed automatically using semicustom design approaches. Only the most performance-critical modules such as the integer and floating-point execution units are handcrafted.

Even though the amount of design automation in the custom design process is minimal, some design tools have proven to be

indispensable. Together with circuit simulators, these programs form the core of every design-automation environment.

Layout Editor

The editor is the premier working tool of the designer and exists primarily for the generation of a physical representation of a design, given a circuit topology. Virtually every design-automation vendor offers an entry in this field. Most well-known is the MAGIC tool developed at the University of California at Berkeley [Ousterhout 84], which has been widely distributed. The basic function of the layout editor is placing polygons on different mask layers so that a functional physical design is obtained (scathingly called polygon pushing).

Since physical design occupies a major fraction of the design time for a new cell or component, techniques to expedite this process have been in continual demand. The symbolic-layout approach has gained popularity over the years. In this design methodology, the designer only draws a shorthand notation for the layout structure. This notation indicates only the relative positioning of the various design components (transistors, contacts, wires). The absolute coordinates of these elements are determined automatically by the editor using a compactor [Hsueh 79, Weste 93]. The compactor translates the design rules into a set of constraints on the component positions, and solves a constrained optimization problem that attempts to minimize the area or another cost function.

An example of a symbolic notation for a circuit topology, called a sticks diagram. The different layout entities are dimensionless, since only positioning is important. You may have noticed that we have made occasional use of such a notation in

earlier chapters. The advantage of this approach is that the designer does not have to worry about design rules, because the compactor ensures that the final layout is physically correct. Thus, she can avoid cumbersome polygon manipulations. Another plus of the symbolic approach is that cells can adjust themselves automatically to the environment. For example, automatic pitch-matching of cells is an attractive feature in module generators. The original cells have different heights, and the terminal positions do not match. Connecting the cells would require extra wiring. The symbolic approach allows the cells to adjust themselves and connect without any overhead.

The disadvantage of the symbolic approach is that the outcome of the compaction phase is often unpredictable. The resulting layout can be less dense than what is obtained with the manual approach. Notwithstanding, symbolic layout tools have improved considerably over the years and are currently a part of the mainstream design process.

Design-Rule Checking

Design rules were introduced in Appendix A as a set of layout restrictions that ensure the manufactured design will operate as desired with no short or open circuits. A prime requirement of the physical layout of a design is that it adheres to these rules. This can be verified with the aid of a design-rule checker (DRC), which uses as inputs the physical layout of a design and a description of the design rules presented in the form of a technology file. Since a complex circuit can contain millions of polygons that must be checked against each other, efficiency is the most important property of a good DRC tool. The verification of a large chip can take hours or days of computation time. One way of

expediting the process is to preserve the design hierarchy at the physical level. For instance, if a cell is used multiple times in a design, it should be checked only once. Besides speeding up the process, the use of hierarchy can make error messages more informative by retaining knowledge of the circuit structure.

DRC tools come in two formats: (1) The on-line DRC runs concurrent with the layout editor and flags design violations during the cell layout. For instance, MAGIG has a built-in design-rule checking facility. (2) Batch DRC is used as a post-design verifier, and is run on a complete chip prior to shipping the mask descriptions to the manufacturer.

4.2.4 Design Synthesis

Synthesis can be defined as the transformation between two different design views. Typically it represents a translation from a behavioral specification of a design entity into a structural description. In laymen's terms, it translates a description of the function a module should perform (the behavior) into a composition, that is, an interconnection of elements (the structure). Synthesis approaches can be defined at each level of abstraction: circuit, logic, and architecture. An overview of the various synthesis levels and their impact is given in Figure 4 - 3. The synthesis procedures may differ dependent upon the targeted implementation style. For instance, logic synthesis translates a logic description given by a state diagram into an interconnection of gates. The techniques involved in this process strongly depend upon the choice of either a two-level (PLA) or a multilevel (standard-cell or gate-array) implementation style. We briefly describe the synthesis tasks at each of the different modeling

levels. We briefly describe the synthesis tasks at each of the different modeling levels. Refer to [DeMicheli 94] for more information and a deeper insight into design synthesis.

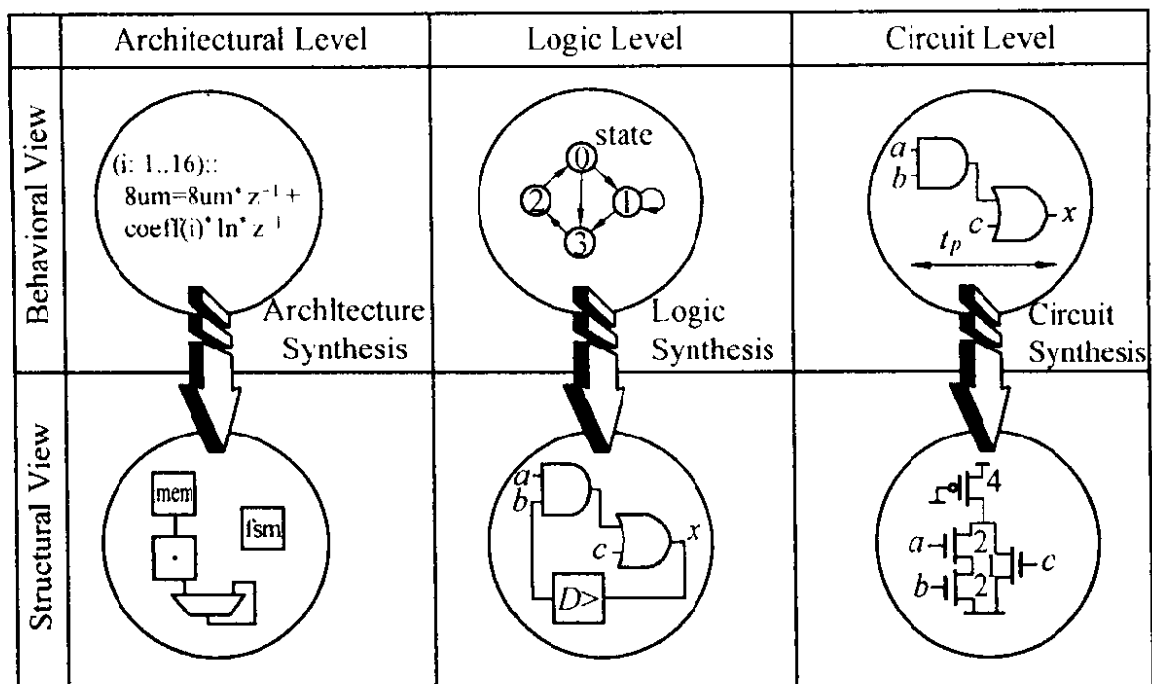


Figure 4 – 3 Switch-level model of CMOS inverter

Circuit Synthesis

The task of circuit synthesis is to translate logic description of a circuit into a network of transistors that meets a set of timing constraints. This process can be divided into two stages.

1. Derivation of transistor schematics from the logic equations. This requires the selection of a circuit style (complementary static, pass-transistors, dynamic, DCVSL, etc.) and the construction of the logic network. The former task is usually up to the designer. The latter depends upon the chosen style. For instance, the logic graph technique introduced in Appendix C can be used to derive the complementary pull-down and pull-up networks of a static CMOS gates. Similarly, automated techniques have been developed to generate the pull-down trees for

the DCVSL logic style so that the number of required transistors is minimized [Chu 86].

2. Transistor sizing to meet performance constraints. This has been a recurring subject throughout this textbook. The choice of the transistor dimensions has a major impact on the area, performance, and power dissipation of a circuit. We have also learned that this is a subtle process. For instance, the performance of a gate is sensitive to a number of layout parasitics, such as the size of the diffusion area, fan-out, and wiring capacitances. Notwithstanding these daunting challenges, some powerful transistor-sizing tools have been developed [e. g., Fishburn 85]. The key to the success of these tools is the accurate modeling of the performance of the circuit using RC equivalent circuits and a detailed knowledge of the subsequent layout-generation process (see the preceding section on compiled cells). The latter allows for an accurate estimation of the values of the parasitic capacitances.

While circuit synthesis has proven to be a powerful tool, it has not penetrated the design world as much as could be expected. One of the main reasons for this is that the quality of the cell library has a strong influence on the complete design and designers are reluctant to pass this important task to automatic tools that might produce inferior results.

logic Synthesis

Logic synthesis is the task of generating a structural view of a logic-level model. This model can be specified in many different ways, such as state transition diagrams, finite state machines, schematic diagrams, Boolean equations, truth tables, or HDL descriptions.

The synthesis techniques differ according to the nature of the circuit (combinational or sequential) or the intended implementation architecture (multilevel logic, PLA, or FPGA). The synthesis process consists of a sequence of optimization steps, the order and nature of which depend on the chosen cost function - area, speed, power, or a combination of the above. Typically, logic optimization systems divide the task into two stages:

1. A technology-independent phase, where the logic is optimized using a number of Boolean or algebraic manipulation techniques.

2. A technology-mapping phase that takes into account the peculiarities and properties of the intended implementation architecture. The technology-independent description resulting from the first phase is translated into a gate netlist or a PLA description.

The two-level minimization tools were the first logic-synthesis techniques to become widely available. The Espresso program developed at the University of California at Berkeley [Brayton 84] is an example of a popular two-level minimization program. For some time, the wide availability of these tools made regular, array-based architectures such as PLAs and PALs the prime choice for the implementation of random logic functions.

At the same time, the groundwork was laid for sequential or state-machine synthesis. Tasks involved include the state minimization that aims at reducing the number of machine states, and the state encoding that assigns a binary encoding to the states of a finite state machine [DeMicheli 94].

The emergence of multilevel logic synthesis environments such as the Berkeley MIS tool [Brayton 87] swung the pendulum

towards the standard-cell and FPGA implementations that offer higher performance or integration density for a majority of random-logic functions.

The combination of these techniques with sequential synthesis has opened the road to complete register-transfer (RTL) synthesis environments that take as an input an HDL description of a sequential circuit and produce a gate netlist. The Synopsys VHDL compiler [Carlson 91] is one of the leaders in this domain and spurred a dramatic change in the way digital circuit design is performed.

Architecture Synthesis

Architecture synthesis is the latest development in the synthesis area. It is also referred to as behavioral or high-level synthesis. Its task is to generate a structural view of an architecture design, given a behavioral description of the task to be executed, and a set of performance, area, and/or power constraints. This corresponds to determining what architectural resources are needed to perform the task (execution units, memories, busses, and controllers), binding the behavioral operations to hardware resources, and determining the execution order of the operations on the produced architecture. In synthesis jargon, these functions are called allocation, assignment, and scheduling [Gajski 92, DeMicheli94]. While these operations represent the core of architecture synthesis, other steps can have a dramatic impact on the quality of the solution. An example of this are optimizing transformations that manipulate the initial behavioral description so that a superior solution can be obtained in terms of area or speed. Pipelining is a typical example of such transformation. In a sense, this component of the synthesis

process is similar to the use of optimizing transformations in software compilers.

While architecture compilers have been extensively researched in the academic community, only recently have they started to make an inroad into the design world, and only in restricted application domains such as digital signal processing. Examples of the latter are the Cathedral-II [DeMan 86] and HYPER [Rebaey 91] synthesis environments. A number of reasons for this slow penetration can be enumerated:

- A lack of understanding of what a behavioral description means at the architecture level for some domains, such as general-purpose microprocessors.

- Behavioral synthesis assumes the availability of an established synthesis approach at the register transfer level. This has only recently come to maturity.

- For a long time, architecture synthesis has concentrated on a limited aspect of the overall design process. For instance, the impact of interconnect on the overall design cost was long ignored. Also, limitations on the architectural scope resulted in inferior solutions apparent to every experienced designer.

4.2.5 Validation and Testing of Manufactured Circuits

While designers tend to spend numerous hours on the analysis, optimization, and layout of their circuits, one issue is often overlooked - when a component returns from the manufacturing plant, how does one know if it actually works? Does it meet the functionality and performance specifications? The customer expects a delivered component to perform as described in the specification sheets. Once a part is shipped or deployed in a

system, it is expensive to discover that it does not work. The later a fault is detected, the higher the cost of correction. For instance, replacing a component in a sold television set means replacement of a complete board as well as the cost of labor. Shipping a nonworking or partially functional device should be avoided if at all possible.

A correct design does not guarantee that the manufactured component will be operational. A number of manufacturing defects can occur during fabrication, either due to faults in the base material (e.g., impurities in the silicon crystal) or as a result of variations in the process, such as misalignment. Other faults might be introduced during the stress tests that are performed after the manufacturing. These tests expose a part to cycles of temperature and mechanical stress to ensure its operation over a wide range of working conditions. Typical faults include short-circuits between wires or layers and broken interconnections. This translates into network nodes that are either shorted to each other or to the supply rails, or that may be floating.

Making sure a delivered part is operating correctly under all possible input conditions is not as simple as it would seem at a first glance. When analyzing the circuit behavior during the design phase, the designer has unlimited access to all the nodes in the network. He is free to apply input patterns and observe the resulting response at any node he desires. This is not the case once the part is manufactured. The only access one has to the circuit is through the input-output pins. A complex component such as a microprocessor is composed of millions of transistors and contains an uncountable number of possible states. It is a very lengthy process to bring such a component into a particular state

and to observe the resulting circuit response through the limited bandwidth offered by the input-output pads - if at all possible. Hardware testing equipment tends to be very expensive and every second a part spends in the tester adds to its price.

It is therefore advisable to consider the testing early in the design process. Some small modifications in a circuit can help make it easier to validate the absence of faults. This approach to design has been dubbed design-for-testability (DFT). While often despised by circuit designers who prefer to concentrate on the exciting aspects of design, such as transistor optimization, DFT is an integral and important part of the design process and should be considered as early as possible in the design flow. "If you don't test it, it won't work! (Guaranteed)" [Weste93]. A DFT strategy contains two components:

1. Provide the necessary circuitry so that the test procedure can be swift and comprehensive.
2. Provide the necessary test patterns (excitation vectors) to be employed during the test procedure. For reasons of cost, it is desirable that the test sequence be as short as possible while covering the majority of possible faults.

In the subsequent sections, we briefly cover some of the most important issues in each of these domains. Before doing so, a short description of a typical test procedure can help to put things in perspective.

Test Procedure

Manufacturing tests fall into a number of categories depending upon the intended goal.

• The diagnostic test is used during the debugging of a chip or board and tries to accomplish the following: Given a failing part,

identify and locate the offending fault.

- The functional test (also called go/no go test) determines whether or not a manufactured component is functional. This problem is simpler than the diagnostic test since the only answer expected is yes or no. Since this test must be executed on every manufactured die and has a direct impact on the cost, it should be as simple and swift as possible.

- The parametric test checks on a number of nondiscrete parameters, such as noise margins, propagation delays, and maximum clock frequencies, under a variety of working conditions, such as temperature and supply voltage. This requires a different set-up from the functional tests that only deal with 0 and 1 signals. Parametric tests generally are subdivided into static (dc) and dynamic (ac) tests.

A typical manufacturing test proceeds as follows. The predefined test patterns are loaded into the tester that provides excitations to the device under test (DUT) and collects the corresponding responses. The test patterns are defined in a test program that describes the waveforms to be applied, voltage levels, clock frequency, and expected response. A probe card, or DUT board, is needed to connect the outputs and inputs of the tester to the corresponding pins on the die or package.

A new part is automatically fed into the tester. The tester executes the test program, applies the sequence of input patterns to the DUT, and compares the obtained response with the expected one. If differences are observed, the part is marked as faulty (e. g., with an ink spot). In case of a probe card, the probes are automatically moved to the next die on the wafer. During the scribing process that divides the wafer into the

individual dies, spotted parts will be automatically discarded. For a packaged part, the tested component is removed from the test board and placed in a good or faulty bin, depending upon the outcome of the test. The whole procedure takes in the range of a few seconds per part, making it possible for a single tester to handle thousands of parts in an hour.

4.2.6 Perspective and Summary

In this chapter, we have briefly scanned the complex world of design methodologies and design automation. In the 1980s, we have witnessed an enormous effort in this domain that has yielded an extensive set of design aids that not only help designers analyze and verify a design, but also optimize and synthesize it and make it testable. These tools and design techniques have had a major impact on the way design is performed today and have made possible the exciting and impressive processors and application-specific circuits to which we have become accustomed. Becoming familiar with these design methodologies is an essential part of the learning experience of the beginning digital designer. We hope this chapter, although compressed, has enticed the reader to further explore the numerous possibilities offered by design automation. We have touched on the following issues in this chapter.

- The simulator is the work-horse of the digital designer. Accuracy can be traded off for execution speed by simplifying the device models and by adapting higher-level data and time abstractions. The following simulation tools are often used by the digital designer circuit, timing, switch, logic, and functional. Choosing the correct simulation level can reduce the response time by orders of magnitude.

- The verification approach avoids the dependency on excitation vectors that is one of the major pitfalls of the simulation approach. Electrical and timing verification tools are now an integral part of digital design methodology. Formal verification has made important strides in recent years, but needs a solution to the complexity problem before it can become a mainstream approach.

- A wide variety of circuit-implementation methodologies have come into existence, ranging from the high-performance, high-density, custom-design approach on one end to the programmable, fast-turnaround FPGA on the other. In between, we find techniques such as standard-cell, module-generation, gate-arrays and sea-of-gates. What approach to select depends upon the application at hand and is a function of the performance, power and cost constraints.

- Design synthesis is making a major inroads in the way design is performed and will continue to grow in importance towards the end of the twentieth century. Logic synthesis is already the dominant implementation technique for most application-specific designs, and architectural synthesis is rapidly gaining acceptance. Manual transistor-level design will be mostly confined to the design of subsystems with stringent requirements for either area, power or speed.

- Testability of a design should be considered in the early phases of the design process. A comprehensive design-for-test approach includes techniques to incorporate controllability and observability into the circuit and the accompanying methods to generate a minimum set of test vectors with sufficient fault coverage.

One final observation: Even with the increasing automation of

the digital circuit design process, new challenges are continuously emerging - challenges that require the profound insight and intuition only offered by a human designer.

4.3 Verification

Verification is not a testbench, nor is it a series testbenches. Verification is a process used to demonstrate the functional correctness of a design. We all perform verification processes throughout our daily lives: balancing a checkbook, tasting a simmering dish, associating landmarks with symbols on a map. These are all verification processes.

In this chapter, I introduce the basic concepts of verification, from its importance and cost, to making sure you are verifying that you are indeed implementing what you want. We look at the differences between various verification approaches as well as the difference between testing and verification. I also show how verification is key to design reuse.

4.3.1 What Is a Testbench

The term “testbench”, in VHDL and Verilog, usually refers to the code used to create a pre-determined input sequence to a design. Then optionally observe the response. It is commonly implemented using VHDL or Verilog, but may also include external data files or C routines.

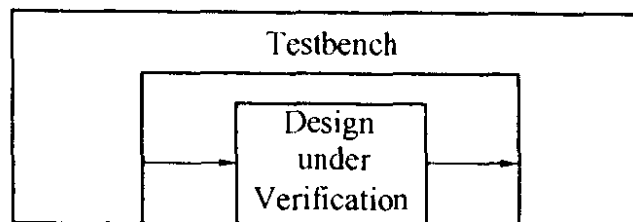


Figure 4 – 7 shows how a testbench interacts with a Design Under Verification

Figure 4 – 7 Generic structure of a testbench and design under test

(DUV). The testbench provides inputs to the design and monitors any outputs. Notice how this is a completely closed system: no inputs or outputs go in or out. The testbench is effectively a model of the universe as far as the design is concerned. The verification challenge is to determine what input patterns to supply to the design and what is the expected output of a properly working design.

4.3.2 The Importance of Verification

If you look at a typical book on Verilog or VHDL, you will find that most of the chapters are devoted to describing the syntax and semantics of the language. You will also invariably find two or three chapters on synthesizable coding style or Register Transfer Level (RTL) subset.

Most often, only a single chapter is dedicated to testbenches. Very little can be adequately explained in one chapter and these explanations are usually very simplistic. In early all cases, these books limit the techniques described to applying simple sequences of vectors in a synchronous fashion. The output is then verified using a waveform viewing tool. Most also take advantage of topic to introduce the file input mechanisms offered by the language, devoting yet more content to detailed syntax and semantics.

Given the significant proportion of literature devoted to writing synthesizable VHDL or Verilog code compared to writing testbenches to verify their functional correctness, you could be tempted to conclude that the former is a more daunting task than the latter. The evidence found in all hardware design teams points to the contrary.

Today, in the era of multi-million gate ASICs, reusable

Intellectual Property (IP), and System-on-a-Chip (SoC) designs, verification consumes about 70% of the design effort. Design teams, properly staffed to address the verification challenge, include engineers dedicated to verification. The number of verification engineers is usually twice the number of RTL designers. When design projects are completed, the code that implements the testbenches makes up to 80% of the total code volume.

Given the amount of effort demanded by verification, the shortage of qualified hardware design and verification engineers, and the quantity of code that must be produced, it is no surprise that, in all projects, verification rests squarely on the critical path. It is also the reason verification is currently the target of new tools and methodologies. These tools and methodologies attempt to reduce the overall verification time by enabling parallelism of effort, higher levels of abstraction and automation.

If effort can be parallelized, additional resources can be applied effectively to reduce the total verification time. For example, digging a hole in the ground can be parallelized by providing more workers armed with shovels. To parallelize the verification effort, it is necessary to be able to write-and debug-testbenches in parallel with each others as well as in parallel with the implementation the design.

Providing higher levels of abstraction enables you to work more efficiently without worrying about low-level details. Using a backhoe to dig the same hole mentioned above is an example of using a higher level of abstraction.

Higher levels of abstraction are usually accompanied by a reduction in control and therefore must be chosen wisely. These

higher levels of abstraction often require additional training to understand the abstraction mechanism and how the desired effect can be produced.

Using a backhoe to dig a hole suffers from the same loss-of-control problem: the worker is no longer directly interacting with the dirt; instead the worker is manipulating levers and pedals. Digging happens much faster, but with lower precision and only by a trained operator. The verification process can use higher levels of abstraction by working at the transaction- or bus-cycle-levels (ore even higher ones), instead of always dealing with low-level zeroes and ones.

Automation lets you do something else while a machine completes a task autonomously, faster, and with predictable results. Automation requires standard processes with well-defined inputs and outputs. Not all processes can be automated. Holes must be dug in a variety of shapes, sizes, depths, locations, and in varying soil conditions, which render general-purpose automation impossible.

Verification faces similar challenges. Because of the variety of functions, interfaces, protocols, and transformations that must be verified, it is not possible to provide a general purpose automation solution for verification given today's technology. It is possible to automate some portion of verification process, especially when applied to a narrow application domain. For example, trenchers have automated digging holes used to lay down conduits or cables at shallow depths. Tools automating various portions of the verification process will be introduced. Hopefully, this book may help define new standard verification practices that could be automated in a near future.

4.3.3 Reconvergence Model

The reconvergence model is a conceptual representation of the verification process. It is used to illustrate what exactly is being verified.

One of the most important questions you must be able to answer is: What are you verifying? The purpose of verification is to ensure that the result of some transformation is as intended or as expected. For example, the purpose of balancing a checkbook is to ensure that all transactions have been recorded accurately and confirm that the balance in the register reflects the amount of available funds.

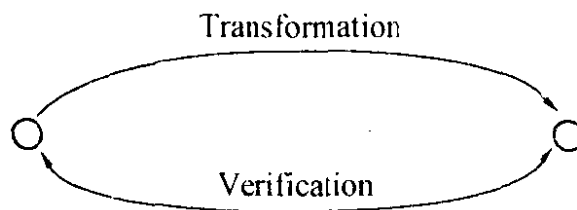


Figure 4 - 8 Reconvergent paths in verification

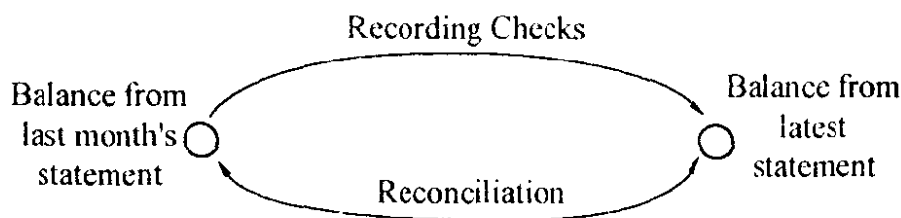


Figure 4 - 9 Balancing a checkbook is a verification process

Figure 4 - 8 shows that verification of a transformation can only be accomplished through a second reconvergent path with a common source. The transformation can be any process that takes an input and produces an output. RTL coding from a specification, insertion of a scan chain, synthesizing RTL code into a gate-level netlist, and layout of a gate-level netlist are some of the

transformations performed in a hardware design project. The verification process reconciles the result with the starting point. If there is no starting point common to the transformation and the verification, no verification takes place.

The reconvergent model can be described using the checkbook example as is illustrated in Figure 4 – 9. The common origin is the previous month's balance in the checking account. The transformation is the writing, recording and debiting of several checks during a one-month period. The verification reconciles the final balance in the checkbook register using this month's bank statement.

4.3.4 The Human Factor

If the transformation process is not completely automated from end to end, it is necessary for an individual (or group of individuals) to interpret a specification of the desired outcome then perform the transformation. RTL coding is an example of this situation. A design team interprets a written specification document and produces what they believe to be functionally correct synthesizable HDL code. Usually, each engineer is left to verify that the code written is indeed functionally correct.

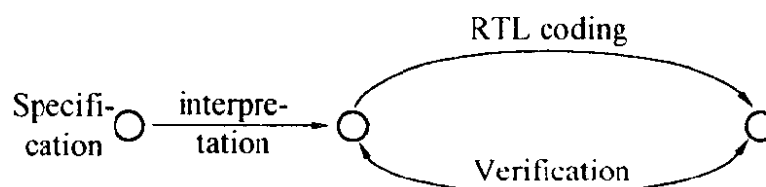


figure 4 – 10 Reconvergent paths in ambiguous situation

Figure 4 – 10 shows the reconvergent path model of the situation described above. If the same individual performs the verification of the RTL coding that initially required interpretation

of a specification, then the common origin is that interpretation, not the specification.

In this situation, the verification effort verifies whether the design accurately represents the implementer's interpretation of that specification. If that interpretation is wrong in any way, then this verification activity will never highlight it.

Any human intervention in a process is a source of uncertainty and unrepeatability. The probability of human-introduced errors in a process can be reduced through several complementary mechanisms: automation, poka-yoka, or redundancy.

Automation

Automation is the obvious way to eliminate human-introduced errors in a process. Automation takes human intervention completely out of the process. However, automation is not always possible, especially in processes that are not well-defined and continue to require human ingenuity and creativity, such as hardware design.

Poka-Yoka

Another possibility is to mistake-proof the human intervention by reducing it to simple, and foolproof steps. Human intervention is needed only to decide on the particular sequence or steps required to obtain the desired results. This mechanism is also known as poka-yoka in Total Quality Management circles. It is usually the last step toward complete automation of a process. However, just like automation, it requires a well-defined process with standard transformation steps. The verification process remains an art that, to this day, does not yield itself to well-defined steps.

Redundancy

The final alternative to removing human errors is redundancy. It is the simplest, but also the most costly mechanism. Redundancy requires every transformation resource to be duplicated. Every transformation accomplished by a human is either independency verified by another individual, or two complete and separate transformations are performed with each outcome compared to verify that both produced the same or equivalent output. This mechanism is used in high-reliability environments, such as airborne systems. It is also used in industries (where later redesign and replacement of a defective product would be more costly than the redundancy itself), such as ASIC design.

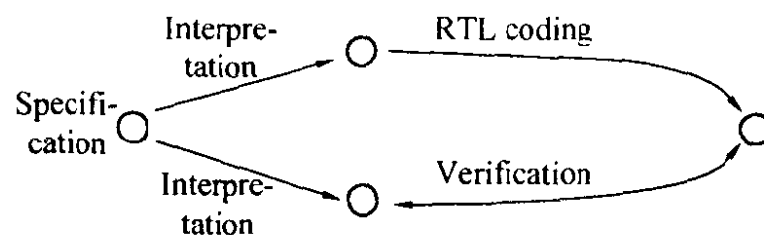


Figure 4 – 11 Redundncy in ambiguous situation

Figure 4 – 11 shows the reconvergent paths model where redundancy is used to guard against misinterpretation of an ambiguous specification document. When used in the context of hardware design, where the transformation process is writing RTL code from a written specification document, this mechanism implies that a different individual must be in charge of the verification.

4.3.5 What Is Being Verified

Choosing the common origin and reconvergence points determines what is being verified. These origin and reconvergence

points are often determined by the tool used to perform the verification. It is important to understand where these points lie to know which transformation is being verified. Formal verification, model checking, functional verification, and testbench generators verify different things because they have different origin and reconvergence points.

Formal Verification

Formal verification is often misunderstood initially. Engineers unfamiliar with the formal verification process often imagine that it is a tool that mathematically determines whether their design is correct, without having to write testbenches. Once you understand what the end point of the formal verification reconvergent paths are, you know what exactly is being verified.

Formal verification falls under two broad categories: equivalence checking and model checking.

Equivalence checking

Figure 4 - 12 shows the reconvergent path model for equivalence checking. This formal verification process mathematically proves that the origin and output are logically equivalent and that the transformation preserved its functionality.

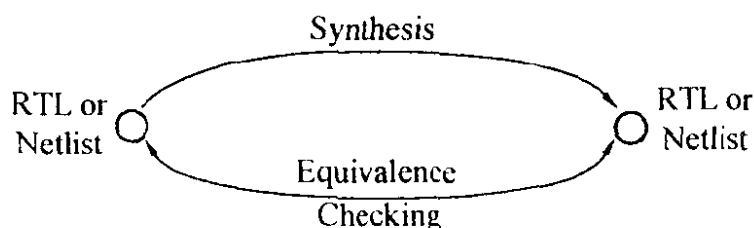


Figure 4 - 12 Equivalence checking compares paths

In its most common use, equivalence checking compares two netlists to ensure that some netlist post-processing, such as scan-chain insertion, clock-tree synthesis, or manual modification, did

not change the functionality of the circuit.

Another popular use is to verify that the netlist correctly implements the original RTL code. If one trusted the synthesis tool completely this verification would not be necessary. However, synthesis tools are large software systems that depend on the correctness of algorithms and library information. History has shown that such systems are prone to error. Equivalence checking is used to verify that manually written RTL code faithfully represents a legacy gate-level design.

Less frequently, equivalence checking is used to verify that two RTL descriptions are logically identical, sometimes to avoid running lengthy regression simulations (when only minor non-functional changes are made to the source code) to obtain better synthesis results.

Equivalence checking is a true alternative path to the logic synthesis transformation being verified. It is only interested in comparing Boolean and sequential logic functions, not mapping these function to a specific technology while meeting stringent design constraints. Engineers using equivalence checking found a design at Digital Equipment Corporation to be synthesized incorrectly. The design used a synthetic operator that was functionally incorrect when handling more than 48 bits. To the synthesis tool's defense, the documentation of the operator clearly stated that correctness was not guaranteed above 48 bits. Since the synthesis tool had no knowledge of documentation, it could not know it was generating invalid logic. Equivalence checking quickly identified a problem that could have been very difficult to detect using gate-level simulation.

Model Checking

Model checking is a more recent application of formal verification technology. In it, assertions or characteristics of a design are formally proven or disproved. For example, all state machines in a design could be checked for unreachable or isolated states. A more powerful model checker may be able to determine if deadlock condition can occur.

Another type of assertion that can be formally verified relates to interfaces. Using a formal description language, assertions about the interface are stated and the tool attempts to prove or disprove them. For example, an assertion might state that, given that signal ALE will be asserted, then either the DTACK or ABORT signal will be asserted eventually.

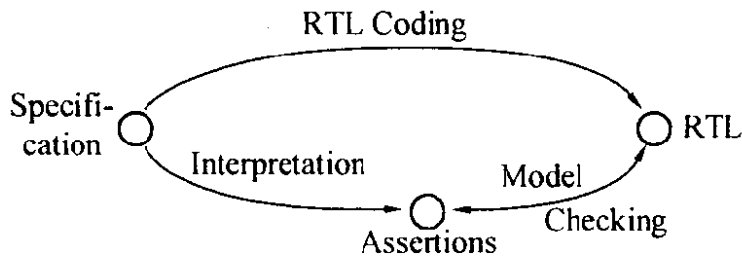


Figure 4 – 13 Model checking paths

The reconvergent path model for model checking is shown in figure 4 – 13. The greatest obstacle for model checking technology is identifying, through interpretation of the design specification, which assertions to prove. Of those assertions, only a subset can feasibly be proven. Current technology cannot prove high-level assertions about a design to ensure that complex functionality is correctly implemented. It would be nice to be able to assert that, given specific register settings, a set of Asynchronous Transfer Mode (ATM) cell will end up at a set of outputs in some relative order. Unfortunately, model checking technology is not at that

level yet.

Functional Verification

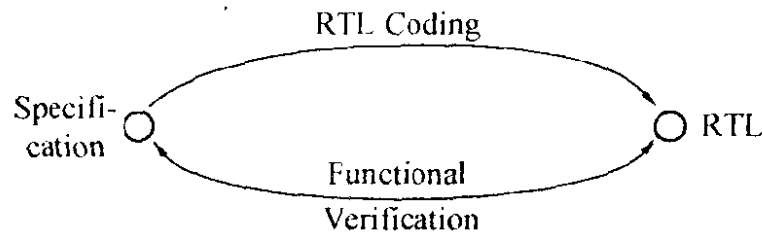


Figure 4 – 14 Functional verification paths

The main purpose of functional verification is to ensure that a design implements intended functionality. As shown by the reconvergent path model in figure 4 – 14, functional verification reconcites a design with its specification. Without functional verification, one must trust that the transformation of a specification document into RTL code was performed correctly, without misinterpretation of the specification's intent.

It is important to note that unless a specification is written in a formal language with precise semantics, it is impossible to prove that a design meets the intent of its specification. Specification documents are written using natural languages by individuals with varying degrees of ability in communicating their intentions. Any document is open to interpretation. Functional verification, as a process, can show that a design meets the intent of its specification. But it cannot prove it. One can easily prove that the design does not implement the intended function by identifying a single discrepancy. The converse, sadly, is not true: no one can prove that there are no discrepancies, just as no one can prove that flying reindeers or UFOs do not exist. (However, producing a single flying reindeer or UFO would be sufficient to prove the opposite!)

Testbench Generation

The increasing popularity of code coverage and model checking tools has created a niche for a new breed of verification tools: testbench generators. Using the code coverage metrics or the results of some proof, and the source code under analysis, testbench generators generate testbenches to either increase code coverage or to exercise the design to violate a property.

These tools appear attractive on the surface, but, as shown in figure 4 – 15, do little to contribute to the verification process. The RTL code is the common origin and there is no reconvergence point. The verification engineer is left to determine if the testbench applies valid stimulus to the design. In the affirmative, he or she must then determine, based on this stimulus, what the expected output is and compare it to the output that was produced by the design.

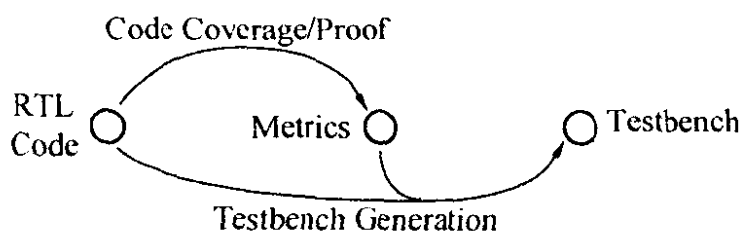


Figure 4 – 15 Testbench generator paths

The usefulness of testbenches generated from code coverage metrics will be discussed in the next chapter. Testbenches generated from model checking results are useful only to illustrate how a property can be violated and what input sequence leads to the improper behavior. It may be a useful vehicle for identification or to provide a debugging environment for fixing the problem.

4.3.6 Functional Verification Approaches

Functional verification can be accomplished using three complementary but different approaches: black-box, white-box, and grey-box.

Black-box verification

With a black-box approach, the functional verification must be performed without any knowledge of the actual implementation of a design. All verification must be accomplished through the available interfaces, without direct access to the internal state of the design, without knowledge of its structure and implementation. This method suffers from an obvious lack of visibility and controllability. It is often difficult to set up an interesting state combination or to isolate some functionality. It is equally difficult to observe the response from the input and locate the source of the problem. This difficulty arises from the frequently long delays between the occurrence of a problem and the apparition of its symptom on the design's outputs.

The advantage of black-box verification is that it does not depend on any specific implementation. Whether the design is implemented in a single ASIC, multiple FPGAs, a circuit board, or entirely in software, is irrelevant. A black-box functional verification approach forms a true conformance verification that can be used to show that a particular design implements the intent of a specification regardless of its implementation.

In very large or complex designs, black-box verification requires some non-functional modifications to provide additional visibility and controllability. Examples include additional or observe internal states, or modify the size of the processed data to

minimize verification time. These registers would not be used during normal operations. They are often valuable during the integration phase of the first prototype systems.

The black-box approach is the only one that can be used if the functional verification is to be implemented in parallel with the implementation of the design itself. Because there is no implementation to know about beforehand, black-box verification is the only possible avenue.

My mother is a veteran of the black-box approach: to prevent us from guessing the contents of our Christmas gifts, she never puts any names on the wrapped boxes. At Christmas, she has to correctly identify the content of each box, without opening it, so it can be given to the intended recipient. She has been known to fail on a few occasions, to the pleasure of the rest of the party!

White-box Verification

As the same suggests, a white-box approach has full visibility and controllability of the internal structure and implementation of the design being verified. This method has the advantage of being able to quickly set up an interesting combination of states and inputs, or isolate a particular function progresses and immediately report any discrepancies from the expected behavior.

However, this approach is tightly integrated with a particular implementation and cannot be used on alternative implementations or future redesigns. It also requires detailed knowledge of the design implementation to know which significant conditions to create and which results to observe.

White-box verification is a useful complement to the black-box verification. This approach can ensure that implementation-specific features behave properly, such as counters rolling over

after reaching their end count value, or datapaths being appropriately steered and sequenced.

Grey-box Verification

Grey-box verification is a compromise between the aloofness of a black-box verification and the dependence on the implementation of white-box verification. The former may not fully exercise all parts of a design, while the latter is not portable.

As in black-box verification, a grey-box approach controls and observes a design entirely through its top-level interfaces. However, the particular verification being accomplished is intended to exercise significant features specific to the implementation. The same verification on a different implementation would be successful but may not be particularly more interesting than any other black-box verification.

4.3.7 Testing Versus Verification

Testing is often confused with verification. The purpose of the former is to verify that the design was manufactured correctly. The purpose of the latter is to ensure that a design meets its functional intent.

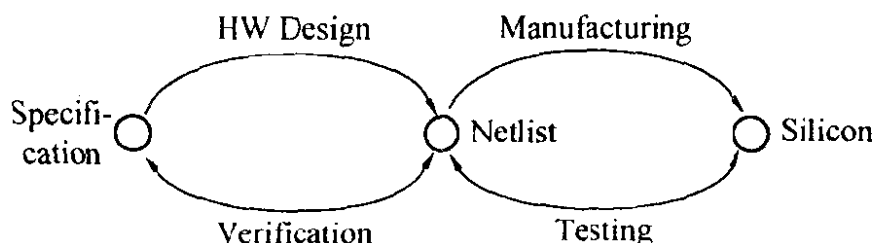


Figure 4 – 16 Test vs. Verification

Figure 4 – 16 shows the reconvergent paths models for both verification and testing. During testing, the finished silicon is reconciled with the netlist that was submitted for manufacturing.

Testing is accomplished through test vectors. The objective of these test vectors is not to exercise functions. It is to exercise physical locations in the design to ensure that they can go from 0 to 1 and from 1 to 0. The ratio of physical locations is called test coverage. The best vectors are usually automatically generated to maximize coverage while minimizing vectors through a process called Automatic Test Pattern Generation (ATPG).

Testing and test coverage depends on the ability to set internal physical locations to either 1 or 0, and then observe that they were indeed appropriately set. Some designs have very few inputs and output, but have a large number of possible states, requiring long sequences to properly observe and control all internal physical locations. A perfect example is an electronic wrist-watch: it has three or four inputs (the digits and symbols on the display). However, if it includes a calendar function, it has billions of possible states (milliseconds in hundreds of years). At speed, it would take hundreds of years to take such a design through all of its possible states.

Scan-based Testing

Fortunately, scan-based testability techniques help reduce this problem to something manageable. With scan-based tests, all registers inside a design are hooked-up in a long serial chain. In normal mode, the registers operate as if the scan chain was not there [see Figure 4 – 17(a)]. In scan mode, the registers operate as a long shift register [see Figure 4 – 17(b)].

To test a scannable design, the unit under test is put into scan mode, then an input pattern is shifted through all of its internal registers. The design is then put into normal mode and a single clock cycle is applied, loading the result of the normal operation

based on the scanned state into the registers. The design is then put into scan mode again. The result is shifted out of the registers (at the same time the next input pattern is shifted in) and the result is compared against the expected value.

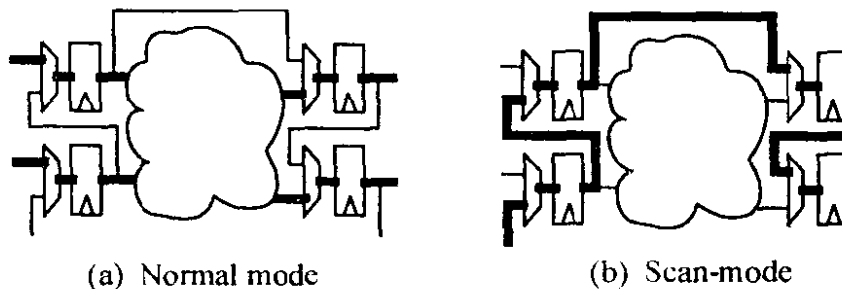


Figure 4 – 17 Scan-based testing

This increase in controllability, and thus test coverage, comes at a cost. Certain restrictions are put onto the design to enable the insertion of a scan chain and the automatic generation of test patterns. Some of these restrictions include, but are not limited to: design fully synchronous design, no derived or gated clocks, and used of a single clock edge. The topic of design for test ability is far greater and complex than this simple introduction implies. For more details, there are several excellent books and papers on the subject.

Hardware designers introduced to scan-based testing initially rebel against the restrictions imposed on them. They see only the immediate area penalty and their favorite design technique rendered inadequate. However, the increased area and additional design effort are quickly outweighed when a design can be fitted with one or more scan chains, when test patterns are generated and high test coverage is achieved automatically, at the push of a button. The time saved and the greater confidence in putting a working product on the market far outweighs the added cost for

scan-based design.

Design for Verification

Design practices need to be modified to accommodate testability requirements. Isn't it acceptable to modify those same design practices to accommodate verification requirements?

With functional verification taking twice as much effort as the design itself, it is reasonable to require additional design effort to simplify verification. Just as scan chains are put in a design to improve testability without adding to the functionality, it should be standard practice to add non-functional structure and features to facilitate verification. This requires that verification be considered at the outset of a project, during its specification phase. Not only should the architect of the design answer the question "what is this supposed to do?", but also "how is this thing going to be verified?"

Typical design-for-verification techniques include providing additional software-accessible registers to control and observe internal locations, and providing programmable multiplexors to isolate or bypass functional units.

4.3.8 Verification and Design Reuse

Today, design reuse is considered the best way to overcome the difference between the number of transistors that can be manufactured on a single chip, and the number of transistors engineers can take advantage of in a reasonable amount of time. This difference is called the productivity gap. Design reuse is a simple concept that seems easy to put in practice. The reality is proving to be more problematic.

Reuse is About Trust

The major obstacle to design reuse is cultural. Engineers have little incentive and willingness to incorporate an unknown design into their own. They do not trust that the other design is as good or as reliable as one designed by themselves. The key to design reuse is gaining that trust.

Trust, like quality, is not something that can be added to a design after the fact. It must be built-in, through the best possible design practices. And it must be earned by standing behind a reusable design: providing support services and building a relationship with the user. Once that trust is established, reuse will happen more often.

Trustworthiness can also be demonstrated through a proper verification process. By showing the user that a design has been thoroughly and meticulously verified, according to the design specification, trust can be built and communicated much faster. Functional verification is the only way to demonstrate that the design meets, or even exceeds, the quality of a similar design that an engineer could do himself or herself.

Verification for Reuse

If you create a design, you have a certain degree of confidence in your own abilities as a designer and implicitly trust its correctness. Functional verification is used only to confirm that opinion and to augment that opinion in areas known to be weak. If you try to reuse a design, you can rely only on the functional verification to build that same level of confidence than custom designs.

Because reusable designs tend to be configurable and programmable to meet a variety of possible environment and applications, it is necessary to verify a reusable design under all

possible configurations and for all possible uses. All claims made about the reusable design must be verified and demonstrated to users.

4.3.9 The Cost of Verification

Verification is a necessary evil. It always takes too long and costs too much. Verification does not generate a profit or make money: after all, it is the design being verified that will be sold and ultimately make money, not the verification. Yet verification is indispensable. To be marketable and create revenues, a design must be functionally correct and provide the benefits that the customer requires.

Verification is a process that is never truly complete. The objective of verification is to ensure that a design is error-free, yet one cannot prove that a design is error-free. Verification can only show the presence of errors, not their absence. Given enough time, an error will be found. The question thus becomes: is the error likely to be severe enough to warrant the effort spent identifying it? As more and more time is spent on verification, fewer and fewer errors are found with a constant effort expenditure. As verification progresses, it has diminishing returns. It costs more and more to find fewer and fewer, often unlikely, errors.

Functional verification is similar to statistical hypothesis testing. The hypothesis under test is: is my design functionally correct? The answer can be either yes or no. But either answer could be wrong. These wrong answers are Type 2 and Type 1 mistakes, respectively.

Figure 4 – 18 shows where each type of mistake occurs. Type

1 mistakes, or false negatives, are the easy ones to identify. The verification is finding an error where none exist. Once the misinterpretation is identified, the implementation of the verification is modified to change the answer from no to yes, and the mistake no longer exists. Type 2 mistakes are the most serious ones: the verification failed to identify an error. In a Type 2 mistake, or false positive situation, a bad design is shipped unknowingly, with all the potential consequences that entails.

	Errors	No Errors
Bad Design		Type 1 (False Positive)
Good Design	Type 2 (False Negative)	

Figure 4 - 18 Type 1 & 2 mistakes

The United States Food and Drug Administration faces Type 2 mistakes on a regular basis with potentially devastating consequences: in spite of positive clinical test result in simple product recall or in the total failure of a space probe after it has landed on another planet.

With the future of the company potentially at stake, the 64-thousand dollar question in verification is: how much is enough? The functional verification process presented in this book, along with some of the tools described in the next chapter attempt to answer that question.

The 64-million dollar question is: when will I be done? Knowing where you are in the verification process, although impossible to establish with certainty, is much easier to estimate than how long it will take to complete the job. The verification

planning process creates a tool that enables a verification manager to better estimate the effort and time required to complete the task at hand, to the degree of certainty required.

4.4 Summary

In this chapter, after briefly outlining what a testbench is, I described the importance of verification. Parallelism, automation and abstractions were identified as strategies to reduce the time necessary to implement testbenches. A model was developed to illustrate and identify what exactly is being verified in any verification methodologies currently available and to differentiate verification from manufacturing test. Techniques for eliminating the uncertainty introduced by human intervention in the verification process were also described. The importance of verification for design reuse and the cost of verification were also discussed.

NOTES

It is only interested in comparing Boolean and sequential logic functions, not mapping these function to a specific technology while meeting stringent design constraints.

仅对布尔逻辑和时序逻辑功能进行比较,而并不用将这些功能映射到具体的工艺技术并满足苛刻的设计约束条件。

Some of these restrictions include, but are not limited to: design fully synchronous design, no derived or gated clocks, and used of a single clock edge.

一些约束包括但并不受限于:完全的同步设计,无派生或门限时钟,使用单边时钟沿。

New Words and Expressions

- layout *n.* 版图
bus *n.* 总线
bypass *n.* 旁路
cell *n.* 单元
propagation delay 传输延迟
excitation vector 激励向量
ripple - carry adder 行波进位加法器
diorama *n.* 立体模型
convergence *v.* 收敛
time queue 时间序列
topology *n.* 布局
hierarchy *n.* 层次
allocation *n.* 分配, 安置
enumerate *v.* 列举
deploy *v.* 配置
die *n.* 芯片
work - horse 重负荷机器
pitfall *n.* 缺陷
testbench *n.* 测试平台
syntax *n.* 语法
semantics *n.* 词义
synchronous *adj.* 同步的
parallelism *n.* 并行化
redundancy *n.* 冗余
discrepancy *n.* 矛盾, 相差
veteran *adj.* 有经验的
test vector 测试向量

lest coverage 最小覆盖

HDL 硬件描述语言

DRC 设计规则检查

PLD 可编程逻辑器

EEPROM 电擦除只读存储器

参 考 文 献

- 1 Jan M Rabaey. Digital Integrated Circuits. Prentic-Hall, Inc. , 1996
- 2 Janick Bergeron. Writing Testbenches: Function Verification of HDL Models. Kluwer Academic Publishers, 2001

Chapter 5

Microelectromechanical Systems (MEMS)

5.1 Introduction

During the past decade, several new fabrication techniques have evolved which helped popularize microelectromechanical systems (MEMS), and numerous novel devices have been reported in diverse areas of engineering and science. One such area is microwave and millimeter wave system. MEMS technology for microwave applications should solve many intriguing problems of high-frequency technology for wireless communications. The recent and dramatic developments of personal communication devices forced the market to acquire miniaturized efficient devices, which is possible only by the development of radio frequency(RF) MEMS.

The term RF MEMS refers to the design and fabrication of MEMS for RF integrated circuits. It should not be interpreted as the traditional MEMS devices operating at RF frequencies. MEMS devices in RF MEMS are used for actuation or adjustment of a separate RF device or component, such as variable capacitors, switches, and filters. Traditional MEMS can be divided into two classes: MEMS actuators and MEMS sensors. The first one is a

kind of moving mechanism activated by an electrical signal like Micromotor. Micro sensors are currently available for a large number of applications. Historically, owing to their ease of fabrication, these were the earliest microsystems. Another reason for the actuators not becoming popular is that the amount of energy generated by such tiny systems does not cause much impact in the associated systems. However, it can be seen later, for microwave and millimeter wave systems, these forces are sufficient to change the properties of overall systems. Passive devices include bulk micromachined transmission lines, filters and couplers. Active MEMS devices include switches, tuners and variable capacitors. The electromotive force used to move the structures on the wafer surface is typically electrostatic attraction, although magnetic, thermal or even gas-based microactuator structures have been developed.

Following the classical review paper by Brown(1998), the RF MEMS development to date can be classified into the following categories based on whether one takes an RF or MEMS view point: (1) RF extrinsic in which the MEMS structure is located outside the RF circuit and actuates or controls other devices in the RF circuit. In this class, one would consider the example of a tunable microstrip transmission line and associated phased shifters and arrays. Microstrip lines are extensively used to interconnect high-speed circuits and components because they can be fabricated by easy automated techniques. (2) RF intrinsic in which the MEMS structure is located inside the RF circuit and has both the actuation and RF-circuit function. In this class, one could consider traditional cantilever and diaphragm type MEMS which can be used as electrostatic microswitch and comb-type capacitors

(Brown, 1998). With the invention of electroactive polymers (EAPs), multifunctional smart polymers and microstereo lithography, these types of RF MEMS can be easily conceived with polymer-based systems. They are also flexible, stable and long lasting. Moreover, they can be integrated with the organic thin film transistor. (3) RF reactive in which the MEMS structure is located inside, where it has an RF function that is coupled to the attenuation. In this class, capacitively coupled tunable filters and resonators provide the necessary RF function in the circuit. Microwave and millimeter wave planar filters on thin dielectric membrane show low loss, and are suitable for low-cost, compact, high-performance mm-wave one-chip integrated circuits.

One of the earliest reported applications of silicon-based RF MEMS technology for microwave applications is in the area of surface micromachined actuators for the realization of microwave switches. These possess very high linearity, low dc standby power and low insertion loss (Larson, 1999). These switches are based on electrostatic attraction counterbalanced by suitable mechanical forces on the beam to pull the switch into the right position. This switch can be designed to present nearly 50Ω impedance across a broad range of frequencies when closed, and nearly an open circuit when there is no connection. This property makes this an attractive choice for microwave applications. Several new switch architectures have also been reported, including the air-bridge structure (Goldsmith, Eshelman and Dennston, 1998). This structure utilizes very high capacitance variation to achieve the switching action. This scheme, however, suffers from relatively high switching voltage requirements.

MEMS technology is also used for RF applications in the area

of variable capacitors, as a replacement for varactor diodes as tuners (Wu et al., 1998). Here, either a lateral or a parallel plate capacitance variation can be obtained with suitable fabrication approaches. The capacitance variation in the parallel plate version is over 3 : 1 making them attractive for wide-band tuning of monolithic voltage-controlled oscillators (VCOs). However their range is often limited by the low-frequency mechanical resonance of the structure.

5.2 MEMS

The term MEMS refers to a collection of microsensors and actuators which can sense its environment and have the ability to react to changes in that environment with the use of a microcircuit control. They include, in addition to the conventional microelectronics packaging, integrating antenna structures for command signals into micro electromechanical structures for desired sensing and actuating functions. The system also may need micropower supply, micro relay and microsignal processing units. Microcomponents make the system faster, more reliable, cheaper and capable of incorporating more complex functions.

In the beginning of the 1990s, MEMS emerged with the aid of the development of integrated circuit (IC) fabrication processes, where sensors, actuators and control functions are co-fabricated in silicon. Since then, remarkable research progresses have been achieved in MEMS under strong capital promotions from both government and industry. In addition to the commercialization of some less-integrated MEMS devices, such as microaccelerometers, inkjet printer heads, micro mirrors for projection, etc., the concepts and feasibility of more complex MEMS devices have been

proposed and demonstrated for the applications in such varied fields as microfluidics, aerospace, biomedicine, chemical analysis, wireless communications, data storage, display, optics, etc. (Fujita, 1996, 1998). Some branches of MEMS, such as micro-opto-electromechanical systems (MOEMS), micro total analysis systems (μ TAS), etc., have attracted a great deal of research interest since their potential application market. As of the end of the 1990s, most MEMS devices with various sensing or actuating mechanisms were fabricated using silicon bulk micromachining, surface micromachining and LIGA¹ processes (Bustillo, Howe and Muller, 1998; Guckel, 1998; Kovacs, Maluf and Petersen, 1998). Three dimensional microfabrication processes incorporating more materials were presented for MEMS recently when some specific application requirements (e. g. biomedical devices) and microactuators with higher output power were called for in MEMS (Fujita, 1996; Guckel, 1998; Ikuta and Hirowatari, 1993; Takagi and Nakajima, 1993; Taylor et al., 1994; Thornell and Johansson, 1998; Varadan and Varadan, 1996; Xia and Whitesides, 1998).

Micromachining has become the fundamental technology for the fabrication of microelectromechanical devices and, in particular, miniaturized sensors and actuators. Silicon micromachining is the most mature of the micromachining technologies and it allows for the fabrication of MEMS that have dimensions in the submillimeter range. It refers to fashioning microscopic mechanical parts out of silicon substrate or on a silicon substrate, making the structures three dimensional and bringing new principles to the designers. Employing materials such as crystalline silicon, polycrystalline silicon and silicon nitride,

etc., a variety of mechanical microstructures including beams, diaphragms, grooves, orifices, springs, gears, suspensions and a great diversity of other complex mechanical structures has been conceived (Bryzek, Peterson and McCulley, 1994; Fan, Tai and Muller, 1987; Middelhoek and Audet, 1989; Peterson, 1982; Varadan, Jiang and Varadan, 2001).

Sometimes many microdevices can also be fabricated using semiconductor processing technologies or stereolithography on the polymeric multifunctional structures. Stereolithography is a poor man's LIGA for fabricating high aspect ratio MEMS devices in UV-curable semi-conducting polymers. With proper doping, a semiconducting polymer structure can be synthesized and using stereo lithography it is now possible to make three-dimensional microstructures of high aspect ratio. Ikuta and Hirowatari (1993) demonstrated that a three-dimensional microstructure of polymers and metal is feasible using a process named IH Process (integrated hardened polymer stereolithography). Using a UV light source, XYZ-stage, shutter, lens and microcomputer, they have shown that microdevices such as springs, venous valves and electrostatic microactuators can be fabricated. In case of difficulty on the polymeric materials, some of these devices can be micromachined in silicon and the system architecture can be obtained by photoforming or hybrid processing (Ikuta and Hirowatari, 1993; Takagi and Nakajima, 1993; Tani and Esashi, 1995; Varadan, 1995; Varadan and Varadan, 1995, 1996). The photoforming or photo fabrication is an optical method such as the stereolithography, photo mask layering process and IH process which involves solidification of photochemical resin by light exposure. Takagi and Nakajima (1993) proposed new concepts of

“combined architecture” and “glue mechanism” using the photoforming process to fabricate complicated structures by combining components, each of them made by its best fabrication process. Batch processing of such hybrid silicon and polymer devices thus seems feasible.

The combined architecture may also result in sheets of smart skin with integrated sensors and actuators at the μm to mm scale. For some applications (say airfoil surface), the smart skin substrate has to be flexible to conform to the airfoil shape and at the same time it has to be compatible with the IC processing for sensor and smart electronics integration. It has been proposed by Carraway (1991) that polyimide is an excellent material for use as the skin because of its flexibility and IC processing compatibility. The control loop between the sensors and actuators employs the multifunctional materials which provide electrical functionality at selected locations using conductive polymers and electrodes that are connected to on-site antennas communicating with a central antenna. A related and difficult problem, and one which has been largely unaddressed, is the method for telemetry of the data. In some applications, stresses and strains to which the structure is subjected may pose a problem for conventional cabling. In others, environmental effects may affect system performance. Advances in ultra flat antenna technology coupled with MEMS sensors/actuators seems to be an efficient solution. The integration of micromachining and microelectronics on one chip results in so-called smart sensors. In smart sensors, small sensor signals are amplified, conditioned and transformed into a standard output format. They may include microcontroller, digital signal processor, application-specific integrated circuit (ASIC), self-

test, self-calibration and bus interface circuits, simplifying their use and making them more accurate and reliable.

The basic MEMS utilize a diaphragm-based, a microbridge-based or a cantilever-based structure. Special processing steps commonly known as micromachining are needed to fabricate these membranes, cantilever beams, resonant structures, etc., which will be discussed later. For a given application, it may be necessary to have integrated MEMS employing one or more of the basic structures. These three structures provide some feasible designs for microsensors and actuators that eventually perform the desired task in most smart structures. However, the main issues with respect to implementing these structures are the choice of materials that are to be used in fabricating these devices and the micromachining technology that may be utilized. To address the first issue, we note that in all of the three structures proposed the sensing and actuation occur as a result of exciting a piezoelectric layer by the application of an electric field. This excitation brings about sensing and actuation in the form of expansion in the diaphragm, or in the free-standing beam in the microbridge structure, or in the cantilever beam. In the former two cases the expansion translates into upward curvature in the diaphragm or in the free-standing beam, hence resulting in a net vertical displacement from the unexcited equilibrium configuration. In the cantilever case, however, and upon the application of electric field, the actuation occurs by a vertical upward movement of the cantilever tip. Evidently, in all three designs the material system structure of the active part (diaphragm, free-standing beam, or cantilever beam) in the microactuator must comprise at least one piezoelectric layer as well as conducting electrodes for the

application of electric field across this layer. Piezoelectric force is used for actuation for many of the applications mentioned above. Micromachining is employed to fabricate the membranes, cantilever beams and resonant structures.

Microsensors and actuators are fabricated using the well-known micromachining techniques in the microelectronics industry. Three-dimensional microactuators in polymer structures can be achieved using stereolithography on UV-curable backbone-type polymers (Ikuta and Hirowatari, 1993; Takagi and Nakajima, 1993; Tani and Esashi, 1995; Varadan, 1995; Varadan and Varadan, 1995, 1996). In the integrated MEMS device, we may use photoforming processing in achieving the combined sensor and actuator architecture as outlined by Takagi and Nakajima (1993). For large actuation, one could use a flex tensional transducer consisting of a piezoelectric diaphragm bridged into a cavity (Chin, Varadan and Varadan, 1994).

Silicon micromachining has been a key factor for the vast progress of MEMS in the past decade. This refers to the fashioning of microscopic mechanical parts out of silicon substrates and more recently other materials. It is used to fabricate such features as clamped beams, membranes, cantilevers, grooves, orifices, springs, gears, suspensions, etc. These can be assembled to create a variety of sensors. Bulk micromachining is the commonly used method but it is being replaced by surface micromachining which offers the attractive possibility of integrating the machined device with microelectronics which can be patterned and assembled on the same wafer. Thus power supply circuitry, signal processing using ASICs can be incorporated. It is the efficiency of creating several such complete packages using existing technology

that makes this an attractive approach.

Table 5 – 1

	Yield Strength (10^{10} dyne/cm ²)	Knoop Hardness (kg/mm ²)	Young's Modulus (10^{12} dyne/cm ²)	Density (gr/cm ³)	Thermal Conductivity (W/cm°C)	Thermal Expansion ($10^{-6}/^{\circ}\text{C}$)
* Diamond	53	7 000	10.35	3.5	20	1.0
* SiC	21	2 480	7.0	3.2	3.5	3.3
* TiC	20	2 470	4.97	4.9	3.3	6.4
* Al ₂ O ₃	15.4	2 100	5.3	4.0	0.5	5.4
* Si ₃ N ₄	14	3 486	3.85	3.1	0.19	0.8
* Iron	12.6	400	1.96	7.8	0.803	12
SiO ₂ (fibers)	8.4	820	0.73	2.5	0.014	0.55
* Si	7.0	850	1.9	2.3	1.57	2.33
Steel	4.2	1500	2.1	7.9	0.97	12
(max. strength)	4.0	485	4.1	19.3	1.78	4.5
W	2.1	660	2.0	7.9	0.329	17.3
Stainless Steel	2.1	275	3.43	10.3	1.38	5.0
Mo	0.17	130	0.70	2.7	2.36	25
Al						

5.3 Mechanical Characteristics of Silicon

Any consideration of mechanical devices made from silicon must certainly take into account the mechanical behavior and properties of single-crystal silicon (SCS). Table I presents a comparative list of its mechanical characteristics. Although SCS is a brittle material, yielding catastrophically (not unlike most oxide-based glasses) rather than deforming plastically (like most metals), it certainly is not as fragile as is often believed. The Young's modulus of silicon (1.9×10^{12} dyne/cm² or 27×10^6 psi), for example, has a value approaching that of stainless steel, nickel, and well above that of quartz and most other borosilicate, sode-lime, and lead-alkali silicate glasses. The Knoop hardness of silicon (850) is close to quartz, just below chromium (935), and

almost twice as high as nickel (557), iron, and most common glasses (530). Silicon single crystals have a tensile yield strength (6.9×10^{10} dyne/cm² or 10^6 psi) which is at least 3 times higher than stainless-steel wire. In practice, tensile stresses routinely encountered in seed crystals during the growth of large SCS boules, for example, can be over 18 000 psi (40-kg boule hanging from a 2-mm-diameter seed crystal). The primary difference is that silicon will yield by fracturing (at room temperature) while metals usually yield by deforming inelastically.

Despite this quantitative evidence, we might have trouble intuitively justifying the conclusion that silicon is a strong mechanical material when compared with everyday laboratory and manufacturing experience. Wafers do break-sometimes without apparent provocation; silicon wafers and parts of wafers may also easily chip. These occurrences are due to several factors which have contributed to the misconception that silicon is mechanically fragile. First, single-crystal silicon is normally obtained in large (5 – 13-cm-diameter) wafers, typically only 10-20 mils (250 to 500 μ m) thick. Even stainless steel of these dimensions is very easy to deform inelastically. Silicon chips with dimensions on the order of 0.6 cm $\times$ 0.6 cm, on the other hand, are relatively rugged under normal handling conditions unless scribed. Second, as a single-crystal material, silicon has a tendency to cleave along crystallographic planes, especially if edge, surface, or bulk imperfections cause stresses to concentrate and orient along cleavage planes. Slip lines and other flaws at the edges of wafers, in fact, are usually responsible for wafer breakage. In recent years, however, the semiconductor industry has attacked this yield problem by contouring the edges of wafers and by regularly

using wafer edge inspection instruments, specifically designed to detect mechanical damage on wafer edges and also to assure that edges are properly contoured to avoid the effects of stress concentration. As a result of these quality control improvements, wafer breakage has been greatly reduced and the intrinsic strength of silicon is closer to being realized in practice during wafer handling. Third, chipping is also a potential problem with brittle materials such as SCS. On whole wafers, chipping occurs for the same qualitative reasons as breaking and the solutions are identical. Individual die, however, are subject to chipping as a result of saw-or scribe-induced edge damage and defects. In extreme cases, or during rough handling, such damage can also cause breakage of or cracks in individual die. Finally, the high-temperature processing and multiple thin-film depositions commonly encountered in the fabrication of IC devices unavoidably result in internal stresses which, when coupled with edge, surface, or bulk imperfections, can cause concentrated stresses and eventual fracture along cleavage planes.

These factors make it clear that although high-quality SCS is intrinsically strong, the apparent strength of a particular mechanical component or device will depend on its crystallographic orientation and geometry, the number and size of surface, edge, and bulk imperfections, and the stresses induced and accumulated during growth, polishing, and subsequent processing. When these considerations have been properly accounted for, we can hope to obtain mechanical components with strengths exceeding that of the highest strength alloy steels.

General rules to be observed in this regard, which will be restated and emphasized in the following sections, can be

formulated as follows:

1) The silicon material should have the lowest possible bulk, surface, and edge crystallographic defect density to minimize potential regions of stress concentration.

2) Components which might be subjected to severe friction, abrasion, or stress should be as small as possible to minimize the total number of crystallographic defects in the mechanical structure. Those devices which are never significantly stressed or worn could be quite large; even then, however, thin silicon wafers should be mechanically supported by some techniques such as anodic bonding to glass—to suppress the shock effects encountered in normal handling and transport.

3) All mechanical processing such as sawing, grinding, scribing, and polishing should be minimized or eliminated. These operations cause edge and surface imperfections which could result in the chipping of edges, and/or internal strains subsequently leading to breakage. Many micromechanical components should preferably be separated from the wafer, for example, by etching rather than by cutting.

4) If conventional sawing, grinding, or other mechanical operations are necessary, the affected surfaces and edges should be etched afterwards to remove the highly damaged regions.

5) Since many of the structures presented below employ anisotropic etching, it often happens that sharp edges and corners are formed. These features can also cause accumulation and concentration of stress damage in certain geometries. The structure may require a subsequent isotropic etch or other smoothing methods to round such corners.

6) Tough, hard, corrosion-resistant, thin-film coatings such

as CVD SiC or Si₃N₄ should be applied to prevent direct mechanical contact to the silicon itself, especially in applications involving high stress and/or abrasion.

7) Low-temperature processing techniques such as highpressure and plasma-assisted oxide growth and CVD depositions, while developed primarily for VLSI fabrication, will be just as important in applications of silicon micromechanics. High-temperature cycling invariably results in high stresses within the wafer due to the differing thermal coefficients of expansion of the various doped and deposited layers. Low-temperature processing will alleviate these thermal mismatch stresses which otherwise might lead to breakage or chipping under severe mechanical conditions.

As suggested by 6) above, many of the structural or mechanical disadvantages of SCS can be alleviated by the deposition of passivating thin films. This aspect of micromechanics imparts a great versatility to the technology. Sputtered quartz, for example, is utilized routinely by industry to passivate IC chips against airborne impurities and mild atmospheric corrosion effects. Recent advances in the CVD deposition (high-temperature pyrolytic and low-temperature RF-enhanced) of SiC [12] have produced thin films of extreme hardness, essentially zero porosity, very high chemical corrosion resistance, and superior wear resistance. Similar films are already used, for example, to protect pump and valve parts for handling corrosive liquids. As seen in Table I, Si₃N₄, an insulator which is routinely employed in IC structures, has a hardness second only to diamond and is sometimes even employed as a high-speed, rolling-contact bearing material. Thin films of silicon nitride will also find important uses in silicon

micromechanical applications.

On the other end of the thin-film passivation spectrum, the gas-condensation technique marketed by Union Carbide for depositing the polymer parylene has been shown to produce virtually pinhole-free, low-porosity, passivating films in a high polymer form which has exceptional point, edge, and hole coverage capability. Parylene has been used, for example, to coat and passivate implantable biomedical sensors and electronic instrumentation. Other techniques have been developed for the deposition of polyimide films which are already used routinely within the semiconductor industry and which also exhibit superior passivating characteristics.

One excellent example of the unique qualities of silicon in the realization of high-reliability mechanical components can be found in the analysis of mechanical fatigue in SCS structures. Since the initiation of fatigue cracks occurs almost exclusively at the surfaces of stressed members, the rate of fatigue depends strongly on surface preparation, morphology, and defect density. In particular, structural components with highly polished surfaces have higher fatigue strengths than those with rough surface finishes. Passivated surfaces of polycrystalline metal alloys (to prevent intergrain diffusion of H_2O) exhibit higher fatigue strengths than unpassivated surfaces, and, for the same reasons, high water vapor content in the atmosphere during fatigue testing will significantly decrease fatigue strength. The mechanism of fatigue, as these effects illustrate, are ultimately dependent on a surface-defect-initiation process. In polycrystalline materials, these surface defects can be inclusions, grain boundaries, or surface irregularities which concentrate local stresses. It is clear that the

high crystalline perfection of SCS together with the extreme smoothness and surface perfection attainable by chemical etching of silicon should yield mechanical structures with intrinsically high fatigue strengths. Even greater strengths of brittle materials can be expected with additional surface treatments. Since hydrostatic pressure has been shown to increase fatigue strengths, any film which places the silicon surface under compression should decrease the initiation probability of fatigue cracks. Si_3N_4 films, for example, tend to be under tension and therefore impart a compressive stress on the underlying silicon surface. Such films may be employed to increase the fatigue strength of SCS mechanical components. In addition, the smoothness, uniformity, and high yield strength of these thin-film amorphous materials should enhance overall component reliability.

A new rotating disk storage technology which has recently been demonstrated by Iwamura et al. not only illustrates some of the unique advantages derived from the use of silicon as a mechanical material but also indicates how well silicon, combined with wear-resistant Si_3N_4 films, can perform in demanding mechanical applications. Data storage was accomplished by an MNOS charge-storage process in which a tungsten carbide probe is placed in direct contact with a 3-in-diameter silicon wafer, rotating at 3600 r/min. The wafer is coated with 2-nm SiO_2 and 49-nm Si_3N_4 , while the carbide probe serves as the top metal electrode. Positive voltage pulses applied to the metal probe as the silicon passes beneath will cause electrons to tunnel through the thin SiO_2 and become trapped in the Si_3N_4 layer. The trapped charge can be detected as a change in capacitance through the same metal probe, thereby allowing the signal to be read. Iwamura et al. wrote and

read back video signals with this device over 10^6 times with signal degradation, at data densities as high as 2×10^6 bits/cm². The key problems encountered during this experiment were associated with wear of the tungsten carbide probe, not of the silicon substrate or the thin nitride layer itself. Sharply pointed probes, after scraping over the Si₃N₄ surface for a short time, were worn down to a 10- μ m by 10- μ m area, thereby increasing the active recording surface per bit and decreasing the achievable bit density. After extended operation, the probe continued to wear while a barely resolvable 1-nm roughness was generated in the hard silicon nitride film. Potential storage densities of 10^9 bits/cm² were projected if appropriate recording probes were available. Contrary to initial impressions, the rapidly rotating, harshly abraded silicon disk is not a major source of problems even in such a severely demanding mechanical application.

5.4 Microfabrications for MEMS

Silicon micromachining has been a key factor for the vast progress of MEMS. Silicon micromachining refers to fashioning microscopic mechanical parts out of a silicon substrate or on a silicon substrate. Silicon micromachining comprises of two technologies: bulk micromachining, in which structures are etched into silicon substrate, and surface micromachining, in which the micromechanical layers are formed from layers and films deposited on the surface.

Bulk micromachining and surface micromachining are the two major micromachining processes of silicon; silicon wafer bonding is usually necessary for silicon microfabrication. LIGA and three-dimensional (3D) microfabrications have been used for high-aspect

ratio and 3D microstructures fabrication for MEMS.

5.4.1 Bulk Micromachining of Silicon

Bulk micromachining technique was developed in 1960s and allows the selective removal of significant amounts of silicon from a substrate to form membranes on one side of a wafer, a variety of trenches, holes, or other structures (Figure 5 – 1). The bulk micromachining technique can be divided into wet etching and dry etching of silicon according to the phase of etchants. Liquid etchants, almost exclusively relying on aqueous chemicals, are referred to as wet etching, while vapor and plasma etchants are referred to as dry etching.

Bulk micromachining is the most mature of the two silicon micromachining technologies. It emerged in the early 1960s and has been used since then in the fabrication of different microstructures. It is utilized in the manufacturing of the majority of commercial devices-almost all pressure sensors and silicon valves and 90% of silicon accelerometers. The term bulk micromachining comes from the fact that this type of micromachining is used to realize micromechanical structures within the bulk of a single-crystal silicon wafer by selectively removing (‘ etching ’) wafer material. The microstructures fabricated using bulk micromachining may cover the thickness range from submicron to full wafer thickness (200 to 500 μm) and the lateral size range from submicron to the lateral dimensions of a full wafer.

For etching such thick silicon substrate, anisotropic wet etchants such as solutions of potassium hydroxide (KOH), ethylenediamine pyrocatechol (EDP), tetramethylammonium

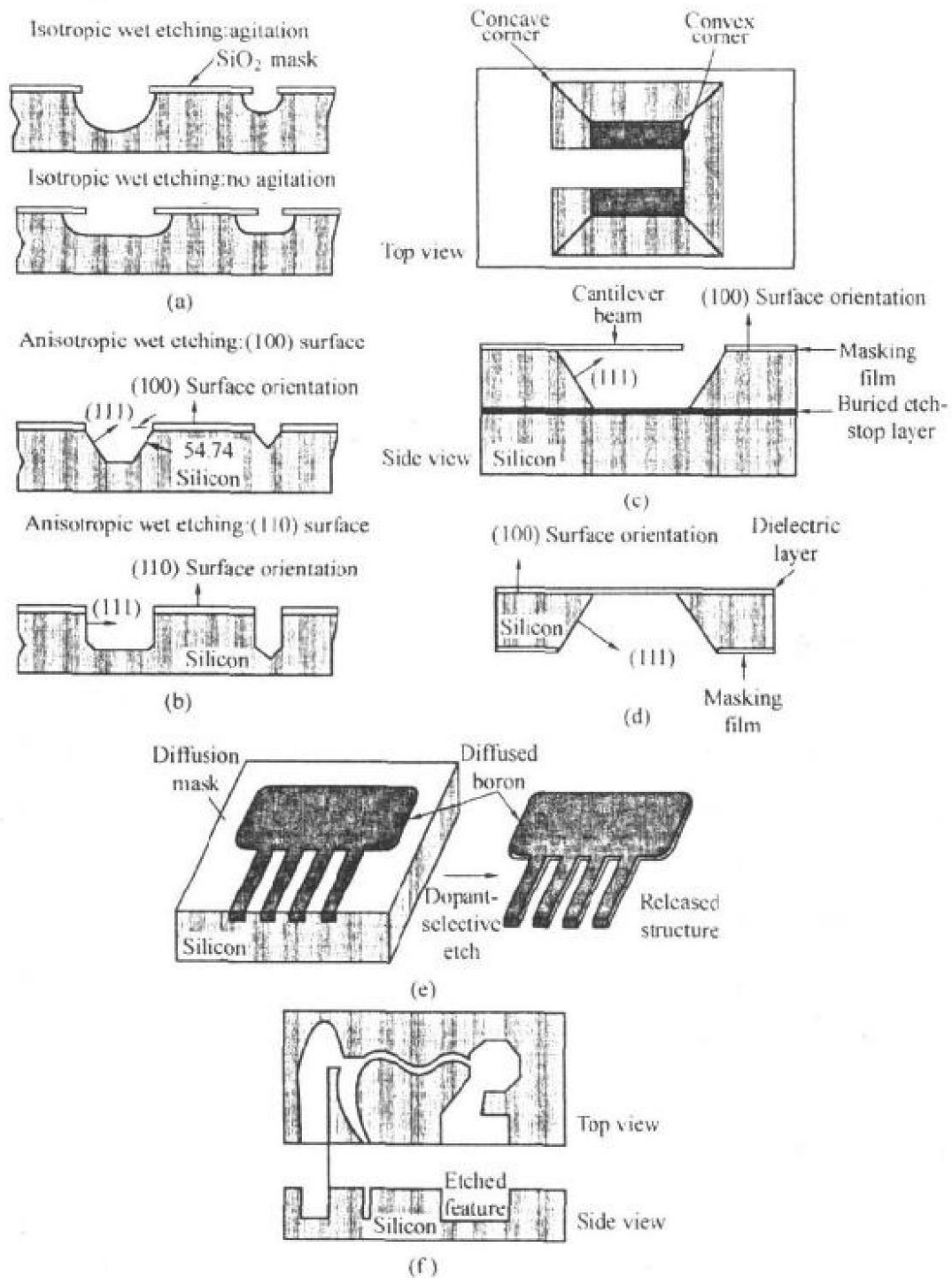


Figure 5 – 1 Bulk silicon micromachining: (a) isotropic etching; (b) anisotropic etching; (c) anisotropic etching with buried etch-stop layer; (d) dielectric membrane released by back-side bulk etching; (e) dopant dependent wet etching. (f) anisotropic dry etching

hydroxide (TMAH) and hydrazine-water are used. These etchants have different etch rates in different crystal orientations of the silicon (Aeidel, 1987; peterson, 1982). Wet etching in most case is done from the back side of the wafer while the plasma-etching is being applied to the front side. In recent years, a vertical-walled bulk micromachining technique known as SCREAM (single-crystal reactive etching and metallization), which is a combination of anisotropic and isotropic plasma etching, is used (Shaw, Zhang and MacDonald, 1994). The etch process can be made selective by the use of dopants (heavily doped regions etch slowly), or may even be halted electrochemically (e. g. etching stops upon encountering a region of different polarity in a biased p-n junction). A region at which wet etching tends to slow down or diminish is called an 'etch-stop'. There are several ways in which an etch-stop region can be created; doping-selective etching (DSE) and bias-dependent DSE (Petersen, 1982; Aeidel, 1982; Shaw, Shang and Macdonald 1994).

Wet etching occurs by dipping substrate into an etching bath or spraying it with etchants which may be acid or alkaline. Wet etching can either be isotropic etching or anisotropic etching depending on the structure of the materials or the etchants used. If the material is amorphous or polycrystalline, wet etching is always isotropic etching [Figure 5 – 1 (a)]. During isotropic etching (etchants used are acid solution), resist is always undercut, meaning the deep etching is not practical for MEMS. Single-crystal silicon can be anisotropically etched. The etching features are determined by the etching speed, which is dependent on the crystal's orientation. The etching slows down significantly at the (111) planes of silicon, relative to other planes. With the chosen

wafers with different crystal orientation, different bulk machined features can be achieved [Figures 5 - 1(b) and 5 - 1(c)]. Most common etchants used for anisotropic etching of silicon include alkali hydroxide etchants (KOH, NaOH, etc.), ammonium-based solutions { NH_4OH , TMAH[$(\text{CH}_3)_4\text{NOH}$], etc.} and EDP (ethylene diamine pyrocatechol, and water). By combining anisotropic etching with boron implantation (P + etch-stop), and electrochemical etch-stop technique, varied silicon microstructures can be bulk machined (Figure 5 - 1).

Dry etching occurs through chemical or physical interaction between the ions in the gas and the atoms of the substrate. Nonplasma, isotropic dry etching can be possible using xenon difluoride or a mixture of interhalogen gases and provides very high selectivity for aluminum, silicon dioxide, silicon nitride, photoresist, etc. The most common dry etching of bulk silicon are plasma etching and reactive ion etching (RIE) etching, where the external energy in the form of RF power drives chemical reactions in low-pressure reaction chambers. A wide variety of chlorofluorocarbon gases, sulfur hexafluoride, bromine compounds and oxygen are commonly used as reactants. The anisotropic dry etching processes are widely used in MEMS because of the geometry flexibility and less chemical contamination than in wet etching sometimes. Arbitrarily oriented features can be etched deep into silicon using anisotropic dry etching (Figure 5 - 1). Very deep silicon Microstructures can be obtained by the deep RIE (DRIE) dry etching (Bryzek, Peterson and McCulley, 1994).

With bulk-micromachined silicon microstructures, the wafer-bonding technique is necessary for the assembled MEMS devices. Surface micromachining, however, can be used to build the

monolithic MEMS devices.

5.4.2 Surface Micromachining of Silicon

Surface micromachining does not shape the bulk silicon but instead builds structures on the surface of the silicon by depositing thin film of 'sacrificial layers' and 'structural layers' and by removing eventually the sacrificial layers to release the mechanical structures (Figure 5 - 2). The dimensions of these surface micromachined structures can be several orders of magnitude smaller than bulk-micromachined structures. The prime advantage of surface-micromachined structures is their easy integration with IC components, since the wafer is also the working for IC elements. It should be noted that as miniaturization is immensely increased by surface micromachining, the small mass structure involved may be insufficient for a number of mechanical sensing and actuation applications.

Surface micromachining requires a compatible set of structural materials, sacrificial materials and chemical etchants. The structural materials must possess the physical and chemical properties that are suitable for the desired application. In addition, they must have satisfactory mechanical properties; e.g. high yield and fracture stresses, minimal creep and fatigue and good wear resistance. The sacrificial materials must have good mechanical properties to avoid device failure during fabrication. These properties include good adhesion and low residual stresses in order to eliminate device failure by delamination and/or cracking. The etchants to remove the sacrificial materials must have excellent etch selectivity and they must be able to etch off the sacrificial materials without affecting the structural ones. In addition the

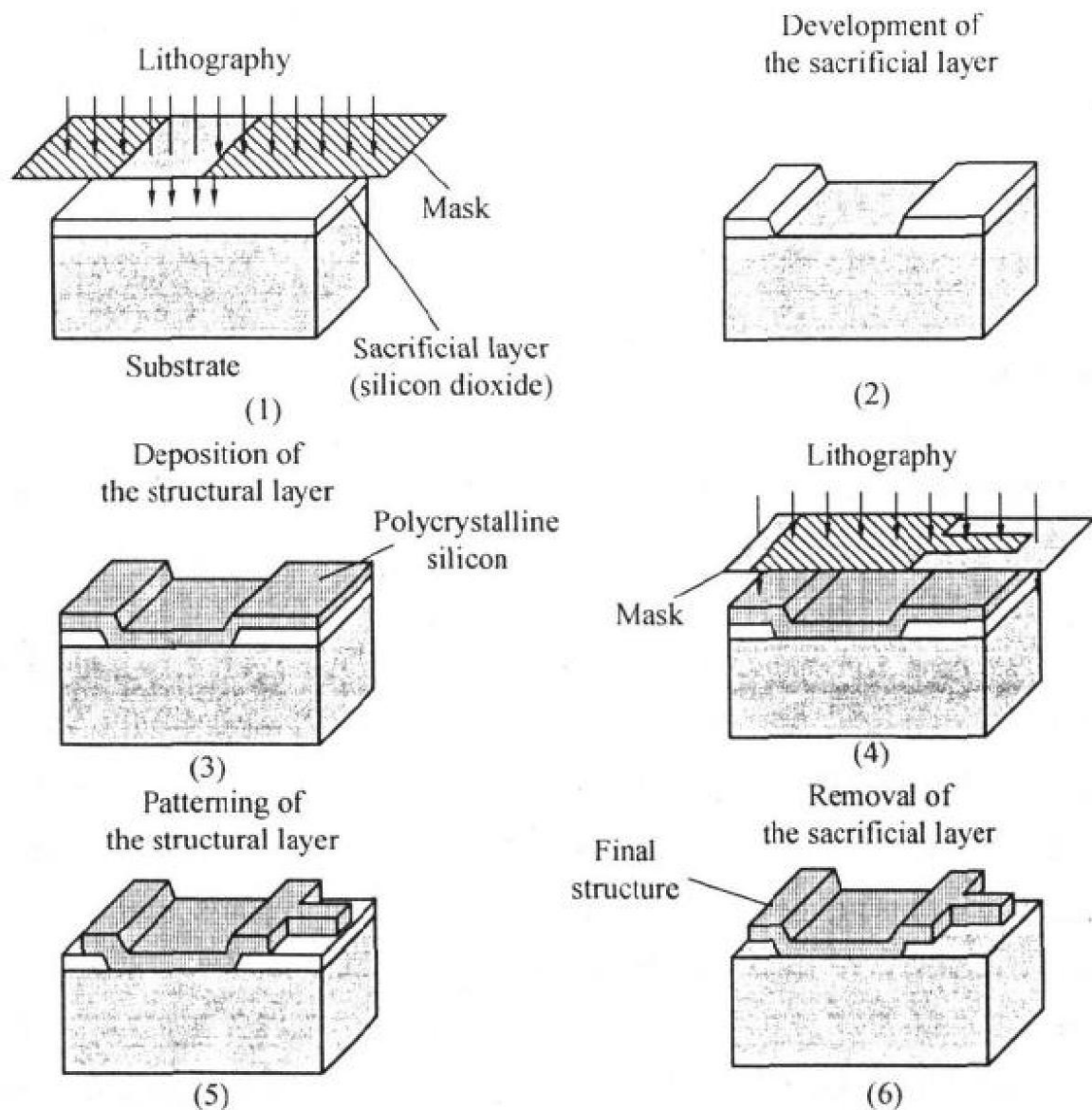


Figure 5 – 2 Processing steps of typical surface micromachining

etchants must have proper viscosity and surface tension characteristics. The common IC compatible materials used in surface micromachining are: (1) polysilicon/Silicon dioxide; low-pressure chemical vapor deposition (LPCVD) deposited polysilicon as the structural material and LPCVD deposited oxide as the sacrificial material. The oxide is readily dissolved in HF solution without the polysilicon being affected. Together with this material system, silicon nitride is often used for electrical insulation. (2) Polyimide/aluminum; in this case polyimide is the structural

material and aluminum is the sacrificial material. Acid-based etchants are used to dissolve the aluminum sacrificial layer. (3) Silicon nitride/polysilicon; silicon nitride is used as the structural material, whereas polysilicon is the sacrificial material. For this material system, silicon anisotropic etchants such as KOH and EDP are used to dissolve polysilicon. (4) Tungsten/silicon dioxide; CVD deposited tungsten is used as the structural material with oxide as the sacrificial material. HF solution is used to remove the sacrificial oxide. Other IC-compatible materials such as silicon carbide, diamond-like carbon, zinc oxide, gold, etc. are also used.

Surface micromachining could also be performed using dry etching methods. Plasma etching of the silicon substrate with SF_6/O_2 -based and CF_4/H_2 -based gas mixtures is advantageous since high selectivities for photoresist, silicon dioxide and aluminum masks can be achieved. However, when using plasma etching, a large undercut of the mask is observed. This is due to the isotropic fluorine atom etching of silicon which is known to be high compared with the vertical etch induced by ion bombardment. In contrast, reactive ion etching of poly-Si using a chlorine/fluorine gas combination produces virtually no undercut and almost vertical etch profiles when using photoresist as a masking material. Thus, rectangular silicon patterns which are up to 30mm deep can be formed using chlorine/fluorine plasmas out of polysilicon films and the silicon wafer surface.

Silicon microstructures fabricated by surface micromachining are usually planar structures (or are two dimensional). Other techniques involving the use of thin-film structural materials released by the removal of an underlying sacrificial layer have

helped to extend conventional surface micromachining into the third dimension. By connecting polysilicon plates to the substrate and to each other with hinges, 3D micromechanical structures can be assembled after release. Another approach to 3D structures used the conformal deposition of polysilicon and sacrificial oxide films to fill deep trenches previously etched in the silicon substrate.

5.4.3 Wafer Bonding for MEMS

Silicon micromachining has limitations in forming complex 3D microstructures in a monolithic format; multichip structures are then proposed for advanced MEMS, where wafer-to-wafer bonding is critical in the formation (Stix, 1992).

The wafer bonding for MEMS can be categorized into three major types: anodic bonding, intermediate-layer assisted bonding and direct bonding.

5.4.3.1 Anodic bonding

Anodic bonding is also called field-assisted thermal bonding, electrostatic bonding, etc. Anodic bonding is usually established between a sodium glass and silicon for MEMS. For the anodic bonding, a cathode and an anode are attached to the glass (or silicon with glass thin coating) and silicon wafer, respectively; voltages applied range from 200 to 1 000 V. At the same time, the anode is put on a heater providing the bonding temperature around 180 to ~ 500 °C (Figure 5 – 3). During the bonding, oxygen ions from the glass migrate into the silicon, resulting in the formation of a silicon dioxide layer between silicon wafer and glass wafer and form a strong and hermetic chemical bond.

The advantage of anodic bonding for MEMS is that the low

temperature used can ensure the metalization layer (aluminum) could withstand this temperature without degradation.

5.4.3.2 Intermediate-layer assisted bonding

This type of bonding for MEMS requires an intermediate layer, which can be metal, polymer, solders, glasses, etc., to fulfill the bonding between wafers (Stix, 1992). One of the earliest wafer bonding-eutectic bonding-utilized gold as the intermediate layer for Si-Si bonding for pressure sensors (Ko, Suminto and Yeh, 1985). The Au-Si eutectic bonding takes place at 363 °C, well below the critical temperature of the metallized aluminium layer. But the stress generated during bonding was found to be significant and introduced sensor drift (Ko, Suminto and Yeh, 1985).

Polymers as an intermediate layer for bonding prevail at very low temperature, reasonable high strength, no metal ion presence and low stress because of the elastic property of polymers, etc. Usually, UV photoresists such as polyimide, AZ-4000, SU-8, polymethylmethacrylate (PMMA), and other UV-curable cross-linked polymers (Madou, 1997). The disadvantage is that the bonded device with polymer may not hold the hermetic sealing performance owing to the relatively high permittivity of polymers.

Glasses with low melting temperature as the intermediate layer for the bonding is also demonstrated, where a layer of glass frit is usually deposited on the silicon wafer. The flatness of the deposited frit layer is critical to obtaining uniform, strong, low-stress bonding. The screen printing of glass frit was used for pressure sensor bonding and exhibits good performance (Ko, Suminto and Yeh, 1985).

Other materials are also being developed as the intermediate

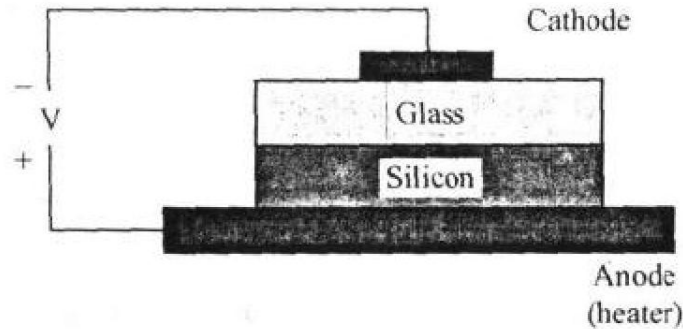


Figure 5 – 3 Anodic bonding

layer for bonding with low temperature, high strength and low stress (Stix, 1992).

5.4.3.3 Direct bonding

Direct bonding is also called silicon fusion bonding, which is used for silicon-silicon bonding. Direct bonding is based on a chemical reaction between OH groups present at the surface of native silicon or grown oxides covering the wafers (Madou, 1997). The direct bonding usually follows three steps: surface preparation, contacting and thermal annealing.

The surface preparation step involves cleaning the surfaces of the two wafers to form a hydrate surface. The wafer surface should be mirror smooth, the roughness should be no greater than $10\text{\AA}$, and the bow of a 4-inch wafer should be less than 5 micron to achieve the necessary flatness (Stix, 1992). Following this preparation, the wafers are aligned and contacted in a cleanroom environment by gently pressing the two wafers at the surface central point. The surface attraction of the two hydrated surfaces then brings the intimate contact over the entire wafer surfaces. The final step in direct bonding is to anneal the bonding from room temperature to $1\,200\text{ }^{\circ}\text{C}$. This annealing process increases the bond strength by more than one order of magnitude at temperatures as

high as 800 to $\sim 1\,200\text{ }^{\circ}\text{C}$. But high-temperature annealing is not allowed for the metallized wafers. The direct bonding prevails in the high-strength bonding, and the devices' dimensions could be scaled down if direct bonding approaches are taken other than anodic bonding.

Some low-temperature direct bonding processes are to be further developed.

5.4.4 LIGA process

MEMS generally require complex microstructures that are thick and three-dimensional (Larson, 1999). Therefore, many microfabrication technologies have been developed to achieve high-aspect-ratio (height-to-width) and 3D devices. The LIGA process is one of those microfabrications.

LIGA is a German acronym for *Lithographie*, *Galvanoformung*, *Abformung* (lithography, galvanofforming, moulding). It was developed by the research Center Karlsruhe in the early 1980s in Germany using X-ray lithography for mask exposure, galvanofforming to form the metallic parts and moulding to produce microparts with plastic, metal, ceramics, or their combinations (Fujita, 1996). A schematic diagram of the LIGA process is shown in Figure 5 - 4. With the LIGA process, microstructures' height can be up to hundreds of microns to millimeter scale, while the lateral resolution is kept at the submicron scale because of the advanced X-ray lithography. Various materials can be incorporated into the LIGA process, allowing electric, magnetic, piezoelectric, optic and insulating properties in sensors and actuators with a high-aspect ratio, which are not possible to make with the silicon-based processes.

Besides, by combining the sacrificial layer technique and LIGA process, advanced MEMS with moveable microstructures can be built. However, the high production cost of LIGA process due to the fact that it is not easy to access X-ray sources limits the application of LIGA. Another disadvantage of the LIGA process relies on that fact that structures fabricated using LIGA are not truly three-dimensional, because the third dimension is always in a straight feature. As we know, complex thick 3D structures are necessary for some advanced MEMS, which means other 3D microfabrication processes need to be developed for MEMS.

5.4.5 Micromachining of Polymeric MEMS Devices

In the micromachining concept for polymeric MEMS devices, two types of polymers are employed: one is the structural polymer and the other one is a sacrificial polymer. The structural polymer is usually a UV-curable polymer with urethane acrylate, epoxy acrylate and acryloxysilane as main ingredients. Its low viscosity allows easy processing through automatic equipment or manual methods without the addition of solvents or heat to reduce the viscosity. It also complies with all volatile organic compound (VOC) regulations. It has excellent flexibility and resistance to fungus, solvents, water and chemicals. Other physical, chemical, mechanical and thermal properties are given in Table 5 – 2 (Varadan, Jiang and Varadan, 2001). This structural polymer may be used as a backbone structure for building the multifunctional polymer described below.

For 3D MEMS devices, the polymers need to have conductive and possibly piezoelectric or ferroelectric properties. For these polymers to be used for polymeric MEMS, they have to meet the

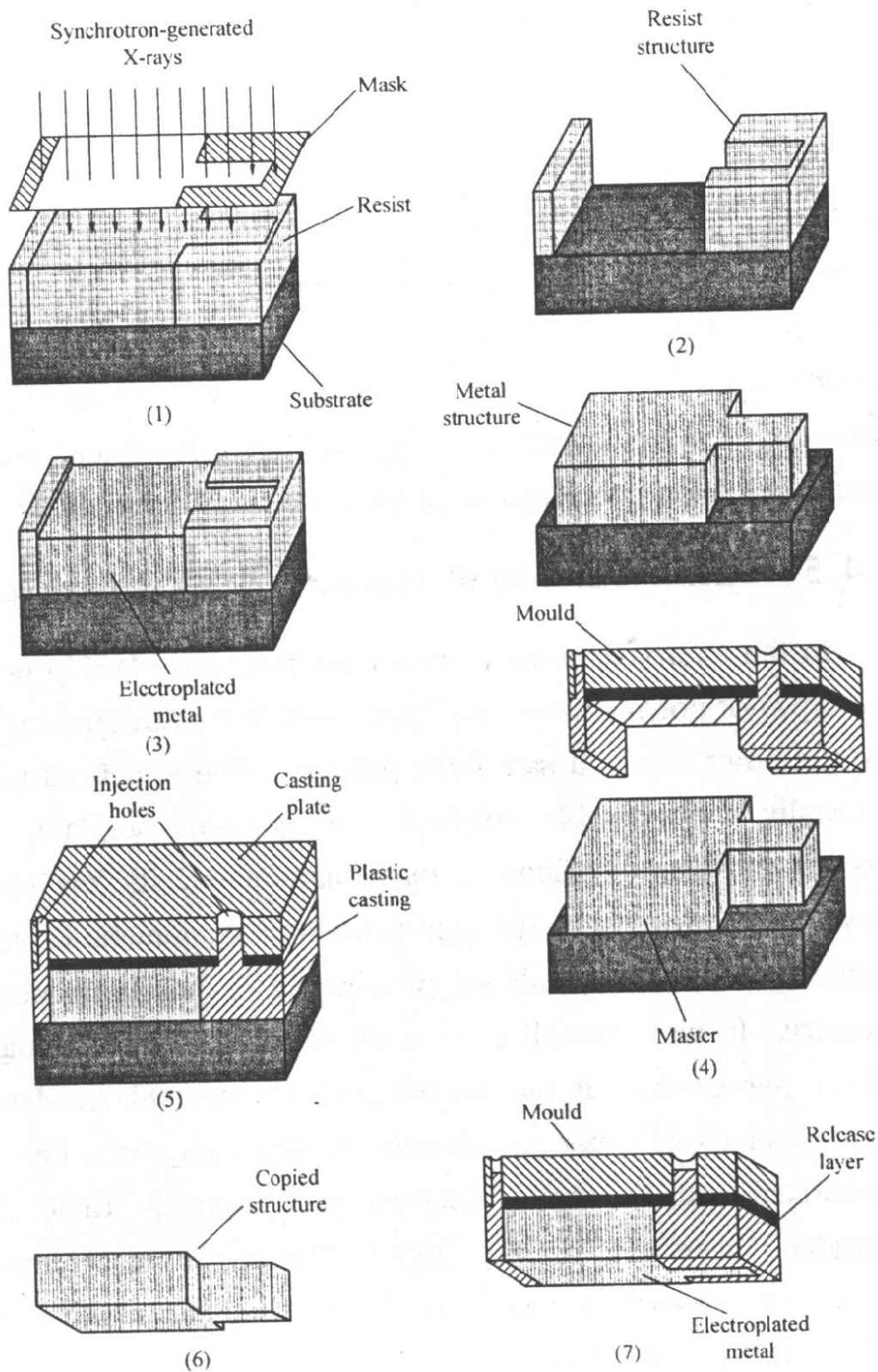


Figure 5 - 4 The LIGA process

following requirements: (1) interactions (chemical or physical)

Table 5 – 2 General properties of polymer

Physical properties:	
clarity	Transparent
flexibility	Good
adhesion(# 600 Cellotape)	Excellent
weather resistance	Excellent
flammability (ASTM D635)	Self-extinguishing
Chemical Properties:	
fungus resistance (ASTM-G21)	Excellent
resistance to solvents	Excellent
resistance to chemicals	Excellent
resistance to water	Excellent
Thermal properties:	
continuous operating range(°C)	65 ~ 125
decomposition temperature(°C)	242
Mechanical properties:	
tensile Strength(psi; ASTM D 683)	3 454
percentage elongation (ASTM D 683)	5.2
Dielectric properties:	
dielectric permittivity (200 ~ 1 000 MHz)	1.9 ~ 2
loss tangent (200 ~ 1 000 MHz)	0.023 ~ 0.05

between functional polymer and nanoceramics; (2) strong interfacial adhesion between functional polymer and conducting polymer layers; (3) suitable elastic moduli to support the deformation initiated by MEMS devices; (4) excellent overall dimensional stability (allowing local mobility); (5) processes conducive to the attachment of nanoceramics and/or conductive phases and formation of a uniform coating layer; (6) long-term environmental stability. In addition, the multifunctionality of these polymers provides a large-scale strain under electric field and thus

can be used as actuators for MEMS-based devices such as micropumps. In general, these polymers are biocompatible and thus useful for many medical devices. Other applications may include implanted medical delivery systems, chemical and biological instruments, fluid delivery in engines, pump coolants and refrigerants for local cooling of electronic components. The sacrificial polymer is an acrylic resin containing 50% silica and is modified by adding Crystal Violet (Varadan, Jiang and Varadan, 2001). This composition is UV curable and can be dissolved with 2 mol⁻¹ caustic soda at 80 °C. In principle this process is similar to the surface micromachining used for silicon devices. However, the process yields 3D structures.

5.4.6 Three-dimensional Microfabrications

To fabricate 3D structures for MEMS, many novel 3D microfabrication techniques have been developed. Among them, microscale freeform fabrications are very impressive to achieve 3D MEMS devices.

Most of the freeform fabrications build 3D microstructures in an additive layer-by-layer fashion (Figure 5 – 5). The members of the freeform microfabrications family include microstereolithography (Ikuta and Hirowatari, 1993), electrochemical fabrication (EFAB; Cohen et al., 1999), microphotoforming (Takagi and Nakajima, 1993), spatial forming (Taylor et al., 1994), microtransfer moulding (Xia and Whitesides, 1998), localized electrochemical deposition (Madden and Hunter, 1996), etc. Complex 3D microstructures have been built using these techniques from the materials of polymer, ceramic, metal, etc. In principle, other smart materials can be

incorporated into these microfabrications.

Another approach to build 3D MEMS devices is to combine the current micromachining processes, such as silicon micromachining, LIGA, precision mechanical machining, etc., or combining the new 3D microfabrication processes with the silicon micromachining and LIGA processes (Bertsch, Lorenz and Renaud, 1998; Takagi and Nakajima, 1994). The AMANDA process is one of these combined microfabrication processes, which combines the LIGA process (or precision machining) with silicon micromachining. AMANDA is the German acronym for *Abformung, Oberflächenmikromechnik und Membranübertragung*. The English translation could be surface micromachining, moulding, and diaphragm transfer. The AMANDA process is specially powerful in polymer MEMS fabrication (Schomburg et al., 1998).

It is known that 3D MEMS devices are usually with multiple layers fabricated from many different structural and functional materials. At the beginning of MEMS fabrications, silicon and polysilicon are the major structural materials, but more structural materials need to be incorporated into MEMS devices since largely divergent applications require varied properties from the microstructures. For example, for the bio-MEMS devices, structural materials should be biocompatible; many polymers are then selected and used to build structures (Ruprecht et al., 1998). While refractory materials may be required for the structures used in high temperature, harsh environments, structural ceramics may be the first choice for this type of application (Ayerdi et al., 1997; Mehregany et al., 1998). Metallic structures have a good reputation in the macro world, as

some of the properties of metals are still maintained in microscale. Therefore, 3D microfabrications capable of processing a broad spectrum of structural materials were expected for MEMS.

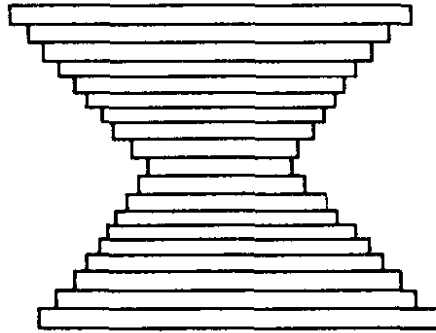


Figure 5 – 5 Freeform three-dimensional microfabrication
(additive)

5.5 Microsensing for MEMS

Various microsensing and microactuation mechanisms have been developed for MEMS for diverse applications. Some of the commonly used sensing and actuating principles are introduced in this section. Many microsensors based on different sensing principles for MEMS have been developed (Fatikow and Rembold, 1997; Rai-Choudhury, 1997), including chemical sensors, gas sensors, optical sensors, biosensors, thermal sensors and mechanical sensors. Some of the major sensing mechanisms for mechanical microsensors are introduced in the following sections.

5.5.1 Piezoresistive Sensing

Piezoresistive sensing utilizes resistors where the resistance is varied through external pressure, to measure such physical parameters as pressure, force and flow rate or to be used as accelerometer.

A typical structure for piezoresistive microsensors is shown in

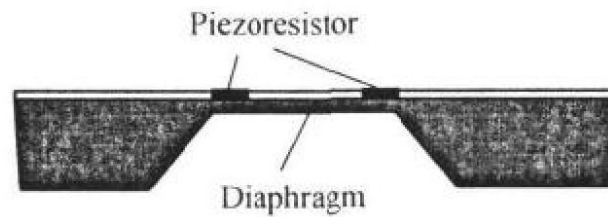


Figure 5 – 6 A piezoresistive sensing structure

Figure 5 – 6. The resistors are usually built on a silicon diaphragm. The deflection of the diaphragm leads to the dimension change of the resistors, resulting in the resistance changing as a result of the piezoresistive effect in silicon:

$$\frac{\Delta R}{R} = (1 + 2\nu) \frac{\Delta l}{l} + \frac{\Delta \rho}{\rho} \quad (5 - 1)$$

where ΔR is the change of the resistance, R is the original resistance, ν is the Poisson ratio, Δl is the length change of the resistor, l is the original length of the resistor, and $\Delta \rho$ and ρ represent the resistivity change and resistors used for this type of piezoresistive microsensor is proportional to the external pressure when the resistivity change is ignored, since the dimension change is proportional to the applied pressure.

Another piezoresistive-type microsensor is schematically shown in Figure 5 – 7, where a semiconductor polymer foil is put on interdigital transducer (IDT) electrodes. If a voltage is applied to the electrodes and there is no pressure applied, the resistance is at the level of mega ohms. When a force is applied, the resistance decreases as a result of current that flows across the shunting polymer foil(Witte and Gu, 1992). Here, the sensing resistance is inversely proportional to the pressure applied.

The performance of the piezoresistive microsensors varies with temperature and the pressure. The sensitivity of the sensors decreases as temperature increases. Any residual stress generated

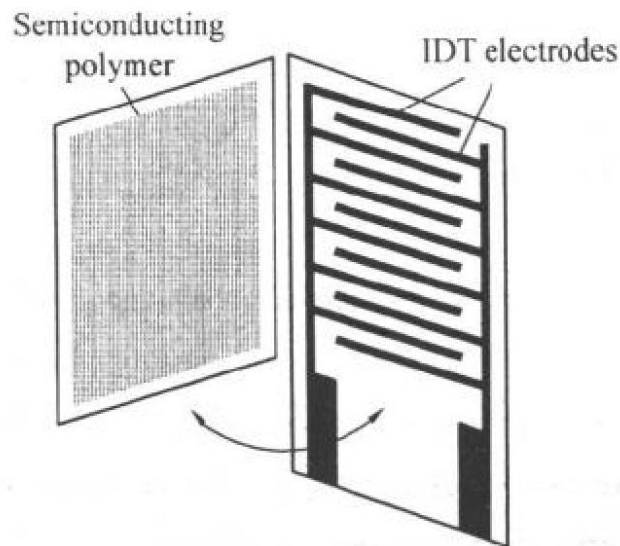


Figure 5 – 7 Piezoresistive sensing combining an interdigital transducer (IDT) and semiconductor polymer

during the fabrication will also influence the sensitivity of the sensors. The nonlinear deflection of the diaphragm occurs when the high-pressure-induced deflection is over 10 % of the diaphragm thickness.

5.5.2 Capacitive Sensing

Capacitive sensing utilizes the diaphragm-deformation-induced capacitance change to convert the information of pressure, force, etc., into electrical signals such as changes of oscillation frequency, time, charge, and voltage. The structure of a typical capacitive microsensor is shown in Figure 5 – 8; an electrode on the flexible diaphragm and the other one on the substrate construct the sensing capacitor. The capacitive microsensors can be used for pressure, force, acceleration, flow rate, displacement, position and orientation measurement, etc.

For capacitive microsensors, the capacitance change is not linear with respect to the diaphragm deformation, and, also, the

small capacitance (generally 1 to ~ 3pF) requires the measurement circuit to be integrated on the chip. But the capacitive sensing was found to have potential for higher performance than piezoresistive sensing in applications requiring high sensitivity, low pressure range and high stability (Rai-Choudhury, 1997).

5.5.3 Piezoelectric Sensing

Piezoelectric sensing is based on the piezoelectric effect of piezoelectric materials. The electrical charge change is generated when a force is applied across the face of a piezoelectric film. For a piezoelectric disc of a given thickness t , the voltage (V) generated across the electrode disc (Figure 5 – 9) when subjected to a stress (T) would be

$$V = gtT \quad (5 - 2)$$

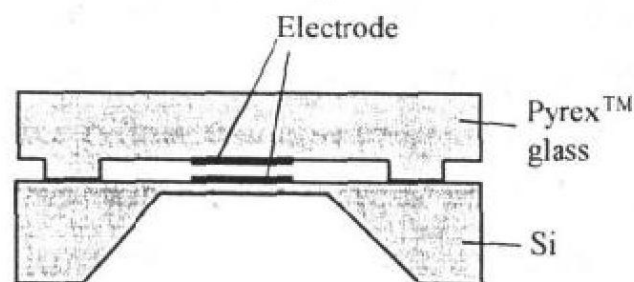


Figure 5.8 Capacitive sensing structure

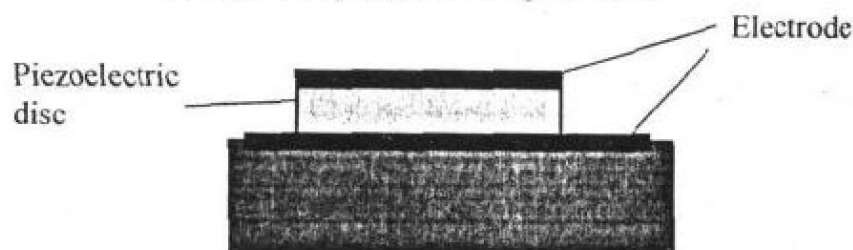


Figure 5 – 8 Piezoelectric sensing

where g is the piezoelectric voltage coefficient. Piezoelectric sensing is mostly used in sensors such as pressure sensors, force sensors, speedometers and accelerometers, hydrophones and

microphones.

5.5.4 Resonant Sensing

Resonant sensing is easily understood as the natural frequency of a string changing as a result of tensile force. In the developed resonant microsensor, strain caused by pressure on the diaphragm leads to the natural frequency of a resonator varying. By picking up the natural frequency variation of the resonator, the physical information that caused the strain will be sensed.

For an example, the natural resonant frequency of a flexure resonator with both ends fixed can be obtained from (Ikeda et al. , 1990)

$$f = \frac{4.73^2 h}{2\pi d^2} \left\{ \frac{E}{12\rho} \left[1 + 0.2366 \left(\frac{l}{h} \right)^2 \epsilon \right]^{1/2} \right\} \quad (5-3)$$

where f is the natural frequency of the fundamental oscillating mode, l is the resonator length, h is the resonator thickness, E is the Young's modulus, ρ is the density of the diaphragm material and ϵ is the strain generated inside the resonator structure. Comparing resonant sensing with piezoresistive sensing, the resonator acts as a kind of strain gauge, the resonant strain gauge, which relates the strain to the resonant frequency. Therefore, the gauge factor of the above resonant strain gauge can be determined as:

$$k_{gf} = \frac{1}{2} \left[0.2366 \left(\frac{l}{h} \right)^2 \right] \left[1 + 0.2366 \left(\frac{l}{h} \right)^2 \epsilon \right]^{-1} \quad (5-4)$$

$$\frac{\Delta f}{f} = k_{gf} \epsilon \quad (5-5)$$

If a strain is 100 ppm, for a 1.2-mm long, 20-micron wide and 5 - micron thick resonator strain gauge, the gauge factor can be as high as 3 000, whereas the piezoresistive strain gauge factor

is only about 2. Since the gauge factor relates directly to the sensitivity of the sensor, the resonant sensing can be used to obtain highly sensitive microsensors. However, resonant sensing usually requires a more complex sensor structure than piezoresistive sensing does; the resonant strain gauges need to be encapsulated from the fluid (Harada et al. , 1999).

5.6 Electromechanical Actuation

A brief comparison of some of the electromechanical transducers discussed above is presented in Table 5 – 3. Owing to its simplicity, electrostatic actuation is the most preferred, especially in microdevices. The control signal here is voltage, which is easy to manipulate in electrical circuits. However, these devices require greater environmental protection as electrostatic fields are prone to attract dust, which could affect the performance of associated CMOS circuits. Electromagnetic and electrodynamic actuators are based on Lorentz force effects. The current-carrying coil is stationary in the former case, whereas in the latter it is moving. These are ideally suited when large currents are possible, even with lower voltages. However, they are prone to problems with power dissipation, but are tolerant to dust and humidity.

Actuators based on piezoelectricity, magnetostriction and electrostriction depend on changes in strain produced by an applied electric or magnetic field in some special materials used. The achievable strain is at a maximum for electrostrictive materials, but the force generated is at a maximum in magnetostrictive materials. Electrostrictive and piezoelectric materials deform with the application of an electric field, but

whereas the relationship between the force produced and applied field is linear in piezoelectrics, it is quadratic in electrostrictive materials.

Most of these transduction schemes are nonlinear. That is, the transfer function between electrical (voltage or current) and mechanical (force or displacement) terms is not linear. Such nonlinearities distort the filtered signal and may cause loss of fidelity. One approach to overcome this difficulty is to restrict the signal to very small variations about a dc bias. It is fairly reasonable to assume that the response to these small signal variations is linear.

Table 5 – 3 Comparison of electromechanical transducers

Actuator	Fractional stroke(%)	Maximum energy density (Jcm ⁻³)	Efficiency	Speed
Electrostatic	32	0.004	High	Fast
Electromagnetic	50	0.025	Low	Fast
Piezoelectric	0.2	0.035	High	Fast
Magnetostrictive	0.2	0.07	Low	Fast
Electrostrictive	4	0.032	High	Fast
Thermal	50	25.5	Low	Slow

Source: Wood, Burdess and Hariss, 1996.

5.7 Materials for MEMS

A broad spectrum of materials has been incorporated into MEMS. In addition to silicon materials, metal, metal alloys, ceramics and polymers are the four major material families used for MEMS(Larson, 1999).

5.7.1 Metal and Metal Alloys for MEMS

Thin-film metals have been used in IC chips for a long time; metal thick-film structures are required for some MEMS devices (Larson, 1999). Microelectroplating and photoforming are used to build thick-film metal structures (Romankiw, 1997; Taylor et al., 1994). Nickel, copper and gold have been electroplated to form thick-film structures, and 3D stainless-steel microparts have been fabricated by photoforming (Taylor et al., 1994). Most of the thick-film metals are used as structural materials of final devices, or as mould inserts for polymer on ceramic micromoulding.

Various metal alloys and the related processes have also been developed for MEMS. CoNiMn thin films were used as permanent magnet materials for magnetic actuation. NiFe permalloy thick films were electroplated onto a silicon substrate for magnetic microelectromechanical system devices, such as micromotors, microactuators, microsensors and integrated power converters, which could allow the production of new micropower magnetics on a chip with integrated circuits (Ahn and Allen, 1998). TiNi shape memory alloy films were sputtered on substrate for SMA sensing and actuating (Ohta et al., 1998). TbFe and SmFe thinfilms were used for magnetostrictive actuation (Honda, Arai and Yamaguchi, 1997).

5.7.2 Polymers for MEMS

Polymers have been extensively used as both structural and functional materials for microdevices. As structural materials, the elasticity, optical properties and biocompatibility of polymers are utilized most in microdevices. Various polymer devices are made

from thin polymer films, thick polymer films and 3D polymer microstructures. Some of the polymers and the related processes for polymer structure fabrication are listed in Table 5 – 4.

It should be pointed out here that the structural polymers listed in Table 5 – 4 can also be used to construct sensing and actuating components for MEMS, just silicon and polymer can be used to build microsensors and actuators, although they act as structural materials, too. Polymer strain gauges and capacitors can serve as sensing elements for piezoresistive and capacitive microsensors (Varadan and Varadan, 1995), and electrostatic polymer microactuators have been developed from polyimide bellow structures (Minami, Morishita and Esashi, 1999). Another important point is that wafer polymer microfabrication processes are being developed for polymer microdevices; the batch fabrication of polymer MEMS is not a concern any more.

Because the limited sensing and actuating mechanisms can be utilized to develop polymer MEMS just from structural polymers, a large number of species of functional polymers has been developed for MEMS during the past several years (Carraway, 1991). Some of the functional polymers used for MEMS are listed in Table 5 – 5.

Functional polymer—solid-powder composites with magnetic and magnetostrictive properties have been developed for microdevices. The polymer-bonded Terfenol D composites showed excellent magnetostrictivity, which can be utilized for microactuation (Ruiz de Angulo, Abell and Harris, 1996). The polyimide-based ferrite magnetic composites were used as polymer magnets for magnetic microactuators (Lagorce and Allen, 1996).

In addition to being used as sensing and actuating materials, polymers have also been used for electronics materials. Polymer

transistors have been developed. Therefore, integration of polymer sensors, actuators and electronics into polymer MEMS will be practical for some special application.

Table 5 – 4 Polymers and fabrication processes for MEMS

Polymer name	Structure	Process	Property utilized	Reference
Polyimide	Thin film	Coating	Elasticity	Ruprecht et al., 1998
Silicon rubber	Thick film	Moulding	Elasticity	Grosjean, Yang and Tai, 1999
Parylene C	Thin film	Coating	Vapor barrier	Grosjean, Yang and Tai, 1999
PMMA	Thick film	LIGA, hot embossing	Elasticity, optical	Becker and Heim, 1999; Guckel, 1998
Polycarbonate	Thick film	Hot embossing	Elasticity, optical transparency	Pan, Lin and Ni, 1999
PDMS	Thick film	Moulding	Elasticity, biomedical compatibility	Armani, Liu and Aluru, 1999
Epoxy resin	Thick film	Casting	Elasticity	Bohm, Olthuis and Bergveld, 1999
Polysulfone	Thick film	Moulding	Excellent mechanical and chemical resistance over broad temperatures (150 °F to 300 °F)	Varadan and Varadan, 1995

Polymer name	Structure	Process	Property utilized	Reference
Acrylate, urethane, epoxy, etc.	Three-dimensional	Microstereo-lithography	Elasticity, optical transparency, UV-curable	Ikuta and Hirowatari, 1993; Shaw, Zhang and Mac Donald, 1994; Takagi and Nakajima, 1993; Tani and Esashi, 1995; Varadan, 1995

Note: LIGA, Lithographie, Galvanoformung, Abformung (lithography, galvanofarming, moulding); PDMS, polydimethylsiloxane; PMMA, polymethylmethacrylate.

Table 5 – 5 Functional polymers for MEMS

	Functional property	Appilcation
PVDF	Piezoelectricity	Sensor, actuator
Poly(pyrrole)	Conductivity	Sensor, actuator, electric connection
Fluorosilicone	Electrostrictivity	Actuator ^a
Silicone	Electrostrictivity	Actuator ^a
Polyurethane	Electrostrictivity	Actuator ^a

^aSource: Pelrine et al., 1997.

Note: PVDF, polyvinylidene fluoride.

5.7.3 Other materials for MEMS

Ceramics are another major species of material used for MEMS. In fact, SiO₂, Si₃N₄ thin films have ever been used for semiconductor devices and silicon MEMS devices. The thick ceramic film and 3D ceramic structures are also necessary for MEMS for special applications. For example, ceramic pressure microsensors have been developed for pressure measurement in the high-temperature

environments (Jennifer and Allen, 1999), silicon carbide MEMS for harsh environments (Chin, Varadan and Varadan, 1994), etc. In addition to these structural ceramics some functional ceramics such as ZnO, PZT, etc., have also been incorporated into MEMS. The fabrication processes developed for ceramic MEMS include screen printing, tape lamination, micromoulding, Sol-Gel and microstereolithography.

参 考 文 献

- 1 Petersen. Silicon as a mechanical material. Proceedings of the IEEE, 1998 (70):5

Chapter 6

Examples of Scientific and Technological Papers

6.1 The Challenges for Physical Limitations in Si Microelectronics

6.1.1 Introduction

For the past four decades, the semiconductor industry has experienced exponential progress, this is unique not only on its development speed, but also on its influence on the society and life of human being. The feature size of semiconductor devices are continued to scale down for higher integrated density, faster circuit speed and lower power dissipation. Currently the $0.25\ \mu\text{m}$ CMOS technology has become the prevailing technology for VLSI applications. Meanwhile, $0.1\ \mu\text{m}$ CMOS and $0.04\ \mu\text{m}$ NMOS devices have recently been fabricated in research laboratories [1], and by the year 2010, the 64 Gb DRAM will be expected to be fabricated with $0.07\ \mu\text{m}$ lithography. Based on such a trend, some key questions on silicon microelectronics' further development are: what are the fundamental limitations of silicon microelectronics when scaling down? Are we approaching to the limits? What can we do facing to the limitations and the limits?

6.1.2 The Challenges for Physical Limitations in Si Microelectronics

Since 1972, the limitations of Si MOSFET scaling down has been researched widely, the predicted minimum feature size of MOSFET changed from $0.25\ \mu\text{m}$ to $0.1\ \mu\text{m}$, but it was broken through every time, now the $0.04\ \mu\text{m}$ has been fabricated successfully. Originally the research of limitations were focused on the limitations of devices scaling down. However, as the microelectronics technology developed into $0.1\ \mu\text{m}$ and beyond, the researches of limitations should be focused on the whole integrated system instead of only on the devices. And the new concepts, new method should be taken facing to the challenges of limitations. The future opportunities for microelectronics to achieve gigascale integration will be governed by a series of theoretical and practical limitation. These include: fundamental physical limits, material limitations, technology limitations, devices limitations, circuits and system limitations.

Fundamental Physical Limits

The fundamental physical limits bases on the physical laws, such as electromagnetic, quantum mechanics, thermodynamics, which are independent of the properties of specific materials. The fundamental limits include that based on electromagnetic dictates that the propagation velocities of a high speed pulse on a long distance interconnection must be less than the speed of light in free space. The uncertainty principle of quantum mechanics requires that during the switch transition duration t_d , the measurable energy should greater than h/t_d , thus the corresponding average power should be greater than h/t_d^2 ,

(where h is Plank's constant). This gives the forbidden of reducing the power of integrated system. Based on the thermodynamics, the thermal energy of a electron gas requires for changing state should be greater than about γkT , ($\gamma \sim 4$), i.e. the power-delay product $P \cdot t_d$ of a switch transition should be greater than about γkT , where k is Boltzmann's constant and T is the absolute temperature. This gives the forbidden of reducing the power-delay product.

The fundamental limit based on the second law of thermodynamics gives the limits on energy consumption of heat generation.

Material Limitations

Currently, the silicon microelectronics technology matures includes: Si (both crystal and poly), SiO_2 , Al, Cu and silicides. The principal properties of these materials are determined by following parameters: relative dielectric constant (ϵ), carrier mobility (μ), carrier saturation velocity (V_s), breakdown field strength (E_c) and thermal conductivity K . The material limitations include:

(1) The primary interconnect material limitations is defined by relative dielectric constant ϵ , i.e. the propagation velocity of a

high speed pulse $v \leq \frac{c}{\sqrt{\epsilon\epsilon_0}}$.

(2) The switching or storage energy limitation is given by

$\frac{1}{2} \epsilon V_0^3 / E_c$, where V_0 is the potential swing.

(3) The limitation on carrier transmission is $\bar{V}_c$.

(4) The ability of heat removal of a devices is determined by K and v_s .

Facing to the material limitations, searching for new materials to substitute for the conventional material is an effective approaching to breakthrough the limitations.

Technology Limitation

Further progresses of Si CMOS technology rely on the supports of the process and techniques, especially the lithograph tools. Now the new optical lithographic tools operation at deep-ultraviolet wavelength (DUV) such as 248 nm are in reduction lines, which used in 0.25 μm technology. The systems employ a wavelength of 193 nm will be used in the next generation of 0.18 μm technology even 0.15 μm technology. However, though using the advanced techniques such as optical proximity correction (OPC), phase-shift masks (PSM), off-axis illumination (OAI) and so on, the resolution of the optical lithography will be enhanced, there are no current expectations that optical lithographic techniques will extend into the sub 0.1 μm regime. Such a limitation results from either ability or the cost. Thus, when CMOS go into the sub 0.1 μm regime, extension beyond today's optical lithography technologies should be developed.

The new lithographic systems based on X-rays, extreme ultraviolet, projection electron-beam or ion projection maybe potential alternatives. Although great steps have been achieved in some systems such as: proximity X-rays steppers developed by IBM, 13 nm extreme ultraviolet (EUV) steppers developed by Intel, AMD and Motorola, and SCALPEL developed by Lucent Technologies, there are many key difficulties to be overcome. For proximity X-rays steppers, the key challenges are $1 \times$ mask fabrication due to the difficulties in the controls of mechanical stress and defects and the control of mask-wafer gap. The 13 nm

EUV steppers maybe the most expected potential lithographic system by industry. However, key challenges to this technology lie in the radiation source, in the multilayer thin film mirror optics and in mask fabrication. SCALPEL uses a scattering mask constructed from a regular wafer that consists of a silicon-nitride membrane with a low atomic number and a patterned layer of tungsten with a high atomic number. An advantage of such scattering mask compared with the absorbing mask used in conventional e-beam projection systems is that it doesn't absorb the unwanted electrons. So it does not heat up, become warped, or distort the images on the wafer. In addition, SCAPEL uses "step-and-scan" mode popular in steppers to exposure the patterns, which improve both the throughput and the practical resolution. But, for SCAPEL, the key challenge is still the throughput like as other e-beam projection systems, even though it has increased a few order of increased in throughput.

For the ion beam, in spite of several advantages over other competitors such that proximity effect are negligible, the exposure field is large enough with minimal distortion, the conventional photoresists can be used. However, more challenges will be met. The main problems are mask heating, ion beam focus, uniformly exposure of photoresists on the wafer.

Challenges of CMOS Scaling Down to Nanometer Regime

Currently, CMOS devices and circuits are the mainstream of Si Microelectronics due to their low power dissipation, feasible integration and scaleable. Generally, the device limitations is considered as the allowable minimum feature of MOSFET.

To achieve minimum channel length both gate oxide thickness t_{ox} and source/drain junction depth x_j should be as small as

possible. The doping profile of the channel region must be optimized at the same time. Meanwhile, when CMOS devices scaling down to nanometer regime, many new effects have to be considered.

(1) Ultra Thin Gate Oxide

During the scaling down of the MOSFET, the gate to channel capacitor must be larger than drain to channel capacitor, in order to keep short channel effect under control and to maintain a good subthreshold turn off slopes, thus gate oxide thickness is reduced nearly in proportion to channel length. For $0.1\ \mu\text{m}$ CMOS devices operating at 1.5 V, the oxide thickness need about 3 nm. This thickness corresponds to roughly ten layers of silicon atoms. With such a thin oxide, quantum mechanical tunneling takes place obviously, leading to gate leakage current that increases exponentially as the oxide thickness is scaled down. At this case, there exist three problems to deal with.

First is tunneling current characteristics itself. Conventional, the transmission probability is used to calculate the tunneling effect by the semi-classic method-WKB approximation or the use of Airy function and numerical integration. But when the oxide is scaling to such thickness, the transmission probability is no longer a meaningful concept and both WKB approximation and Airy function are not valid. New method must be developed, such as by use of quantum mechanical method cupped with a transverse-resonance method.

Second is the influence of tunneling effect on the properties of MOSFET. It can have influence on leakage current, threshold voltage, mobility as well as reliability due to the tunneling effect causing inversion layer charge loss and additional increased

electron fluency passing through the gate oxide. Commonly, tunneling effect was considered as fundamental physical limitation of MOSFET scaling down, and the gate oxide thickness limits was regarded as about 3nm due to the gate leakage current. However, with development of technology, the MOSFET with 1.5nm thickness of SiO₂ gate oxide was developed with ideal uniformity and reliability. Thus the influence of tunneling effect should be researched thoroughly with new concepts and new methods.

The third work is search for the new high dielectric constant materials such as Zr_{0.15}Sn_{0.3}Ti_{0.55}O_{2-δ}, Ta₂O₅-TiO₂, (Sr, Ba) TiO₃ to substitute for SiO₂.

(2) Effect of Dopant Number Fluctuation

As MOSET scaling into the sub 0.1μm regime, the number of dopants is of the order of hundreds in the depletion region for minimum geometry device. As a result, the detailed microscopic dopant distribution in the MOSFET channel will have nonegligible effect on device threshold voltage. Thus, the performance fluctuations of MOSFET can be caused by the statistical fluctuation of the channel dopant number. This problem is considered as to be none of the most serious concerns in gigascale integrated circuit.

(3) High Field Effects in Silicon

Because the power supply has not been scaled in proportion to the channel length, the electric field strength has been increased as CMOS scaling down. For 0.1 μm channel length devices, operating at 1 volt, the electric field in oxide may reach a maximum of 5 MV/cm, while the field in silicon has exceeded 1 MV/cm. They are expected to go even higher when the channel length is scaled into nanometer regime. At such high fields, several undesirable effects occur, which either increase the

leakage current or detract from the device performance. High field effects in gate oxide was discussed in section II(1), while the high fields effects in silicon including: quantum effect in inversion charge layer, band to band tunneling and mobility degradation, which will be discussed following.

(a) Quantum effect in inversion charge layer

The continued scaling of MOSFET gate length requires decreased gate oxide thickness (t_{ox}) and increased channel doping in order to improve device performance. In deep submicron regime, the gate oxide is less than 10 nm, and in 0.1 μm MOSFET, the thickness of gate oxide is only 4 to 3 nm. This results in the sufficiently large transverse electric field at the Si/SiO₂ interface near threshold region and quantization of electrons motion perpendicular to the surface which leads to a splitting of the energy levels into subbands, such that the lowest of the allowed energy levels for electrons in the well do not coincide with the bottom of the conduction band. This phenomenon begin to have an observable influence on room temperature deep submicron MOSFET's performances especially on threshold voltage and C-V characteristics.

(b) band to band tunneling in Si

Usually there exists a high field region near the drain, where a band bending of 1-2 V occurs within a distance as little as 10nm. Under these conditions quantum mechanical band to band tunneling will take place, which could result in significant junction leakage currents.

(c). Mobility degradation

It is well known that both electron and hole mobility's exhibit a "universal" behavior versus the effective field ϵ experienced by

an average carrier. The research works show, when ϵ small than 5×10^5 V/cm, the mobility essentially follows a $\epsilon^{-0.3}$ dependence as the carrier transport is dominated by phonon scattering. At high fields, the carrier are pressed toward Si/SiO₂ interface and surface scattering becomes dominant. Surface scattering is sensitive to the conditions of Si/SiO₂ interface and results in ϵ^{-1} (hole) and ϵ^2 (electron) dependence for the mobility. Because of this stronger dependence, both electron and hole mobilities decrease rapidly when $\epsilon > 10^6$ V/cm. This may degrade the performance gained from scaling down.

Limitation of Circuit and System

(1) Interconnection

As the circuit speeds become faster and chips become larger and more complex due to the shrink in device size and increase in integration level, the interconnection problems become more important and even become a limitation factor for the system performances. The limitation of interconnection results from propagation delay, delay variation, and noise levels which include cross talk and delta-I noise. The characteristics of interconnection delay and noise mainly influenced by the structure and the material of interconnection. Thus, the improvements of interconnection characteristics are mainly dependent on both materials and structures.

New materials with low dielectric constant and low dielectric losses and new conducting wiring with low resistivity should be applied. In structure, the interconnect should be designed optimally, therefore, proper modeling and simulation techniques which can accurately model and simulate the characteristics of interconnections have to be used. In the modeling and simulation

techniques, the interconnect wiring have to be treated as transmission line structure, and nonuniform, discontinuities, the frequency dependence of distribution parameters should be considered, and the short wiring, medium wiring as well as long wiring should be distinguished. In addition, shrinking device sizes is still meaningful even though without gaining the performance from scaling down because the interconnection delay of chip-to-chip is larger than that of on chip.

(2) Reliability

With the feature size of devices scaling down especially the thickness of gate oxide reduced continually, the reliability of devices and the lifetime of sub $0.1\ \mu\text{m}$ and beyond, hot carrier will be somewhat reduced by lowering V_{DD} , and the hot carrier effect is improved by the reduction of V_{DD} . But the new effects which influence on the device reliability such as tunneling effect are not characterize well yet. Metallization (interconnect) reliability is also important for we need and concern is whole integrated system with $10^9 - 10^{10}$ transistors on a chip, not a single transistor. Thus the reliability including ultra thin gate oxide, devices, metallization, as well as whole chip are need to be well characterized by new concepts and new methods.

(3) Heat Removeing

In integrated system with $10^9 \sim 10^{10}$ transistors on a chip, heat removing is also a challenge, much more attention should be paid when designed integrated system.

6.1.3 Next Generation Electronics-Quanium Electrics Devices

In order to continue the miniaturization of circuit and device

down to the nanometer scale even to approach the quantum feature length, researchers are investigating several alternatives to the transistors for ultradense circuits. These new nanometer scale electronic devices perform as both switches and amplifiers, just like today's transistors. However, unlike today's FET's, which operate based on the movement of masses of quantum mechanical phenomena that emerge on the nanometer scale, including the discreteness of electrons.

What will such alternative next generation nano devices look like is still not well understand. Here will introduce some novel nano devices especially solid state quantum effect devices that might replace the transistors in tomorrow's smaller, denser, faster electronics system by surveying the literature.

The nano devices take advantage of various quantum effect can classified as solid-state quantum effect devices and molecular electronic devices.

A number of nanometer scale solid state devices have been suggested, their essential structure feature that all of these devices have in common is a small "island" composed of semiconductor or metal in which electrons maybe confined. The composition, shape, and size of the island gives the different types of solid state nanoelectronic devices and their distinct properties. Controlling these factors permits the designer of the device to employ quantum effects in different ways to control the passage of electrons on the end of the island. There can define three basic categories of solid state nanoelectronic devices: quantum dots (QD), resonant tunneling devices (including resonant tunneling diode, RTD and resonant tunneling transistor, RTT), single electron devices (including single electron transistor, SET and single electron

memory, SEM).

Even through, the considerable progresses have been made, a number of challenging issues exist in solid state quantum effect devices, such as: valley current, sensitivity to input voltage and current fluctuations, cryogenic operation, materials, background charge problem, extreme difficulty of making islands and tunnel barriers precisely and uniformly in solid state devices, and so on.

Molecular electronics uses primarily covalently bonded molecular structure, electrically isolated from a bulk substrate. Devices of this description, wires and switches composed of individual molecules and nanometer scale superamolecular structures. Some molecular electronic switching devices have been proposed as follows: electric-field controlled molecular electronic switching devices, electromechanical molecular electronic devices, electrochemical molecular devices and so on.

New Words and Expressions

optical proximity correction(OPC) 光学逼近校正

phase-shift masks(PSM) 相移掩膜

off-axis illumination(OAI) 离轴照明

tungsten *n.* 钨

cryogenic *adj.* 低温学的

6.2 Microelectronics and Photonics – the Future

We address some major trends in microelectronics photonics, and some other areas of “high-technology” engineering. We discuss which of the new directions and emerging technologies in this field of engineering look most promising, and what challenges

a mechanical/materials/reliability engineer will most likely encounter in connection with these trends and technologies. © 2000 Published by Elsevier Science Ltd.

6.2.1 Introduction

The emergence of and progress in, microelectronics and photonics have revolutionized the telecommunications and information science and engineering in the 20th century. It would be difficult to identify or even conceive of any other contemporary area of engineering that has had a more dramatic, pervasive and beneficial influence on the way we live and work. The extent and sophistication of the spectacular scientific and technological advances that are at our fingertips today as a result of the successes in microelectronics, optoelectronics and photonics are amazing. These “high-technology” areas control everything in today’s world—from a space shuttle to a washing machine — and are equally important from the perspective of an air force pilot, a young video enthusiast, a banking executive, an entertainer or a homemaker. A good example is the astonishing growth of the Internet, with the number of users continuing to double every few months. Microelectronics and photonics have tremendous implications for industry, employment, strategic position of the country, and even for the future organization of society. The well being of the industrial nations depends on the development of intelligent products, tools and processes. It is natural that everyone, not specialists only, wonders which new applications of microelectronics and photonics are most likely to come into life in the near future and what difference these applications might make for the 21st century home and office.

In this overview we discuss some major trends in microelectronics, optoelectronics, photonics and some other areas of “high-technology” engineering, primarily from the viewpoint of a reliability, mechanical and materials engineer. We try to anticipate which of the new directions and emerging technologies look most promising, and what challenges a reliability/mechanical/materials engineer will most likely encounter in connection with these technologies. We would like to point out that to be a “prophet” in predicting the future of many of the “high-technologies” is not as hard as it may seem at first glance: unless something absolutely unexpected happens, the current trends indeed determine what will take place in the foreseeable future. This is due to the tremendous momentum of these trends that is supported by the existing knowledge, powerful customer demand and the enormous investments of capital and human resources.

6.2.2 Microelectronics

6.2.2.1 History and trends

The transistor invented in 1947 at Bell Laboratories and the integrated circuit (IC) built independently at Texas Instruments and Fairchild Company 12 years later revolutionized the electronics industry. Since then, the major trends in microelectronics were the drives for smaller, faster, more reliable and less expensive ICs. Exponentially growing technological, productivity and cost improvements enabled expanding markets and revenues, which, in turn, enabled larger R&D and manufacturing investments. These, in turn, enabled again technological, productivity and cost improvements. This paradigm has served well for more than four

decades. The semiconductor industry is probably unique in maintaining such a steep performance/cost improvement trend for its products over such a long period of time.

The progress in microelectronics is measured, first of all, by the degree of miniaturization, i. e. by the number of devices that could be crammed onto a silicon chip. The desire to make the devices smaller and smaller brought to life several generations of the IC technology: small scale integration (SSI), medium scale integration (MSI), large scale integration (LSI), and very large scale integration (VLSI). We live now in the VLSI era, when a typical chip contains about ten million devices. The existing technologies are incapable yet of producing ICs on a silicon wafer with 100% yield. If it happens, then an era of ultra large scale integration (ULSI), or wafer scale integration (WSI), will commence.

The number of electronic devices on a chip is still doubling approximately every 18 months. This trend will probably have matured by 2010. By that time highly packed ICs (such as memory chips) will contain up to a few billion devices, and the device density for custom logic ICs will increase from about 100 million to approximately one billion devices per chip. Such a trend will be, in part, accelerated by the fact that chips will grow in size. This, in combination with the trend towards thinner chips, will provide an obvious challenge for a reliability engineer. Another challenge will be due to the fact that chips with a larger number of devices, which will have substantially higher frequencies (speeds of 200 MHz and higher are becoming more and more common), will dissipate more power. The effective handling of the increased chip size and elevated heat dissipation relies on the continuous effort in

improving designs, materials and manufacturing technologies.

Silicon wafers will also grow in size. A typical wafer today is 4 in in diameter, and the largest wafers are twice as large. The industry is moving to 12 or even to 16 in (a Japanese program) wafers. It is expected that 12 in wafers will be introduced into high volume production by the year 2010. Mini-batch processing and, perhaps, single-wafer technologies will play an important role in the 12 in wafer era.

There is a crucial need for a continuous effort in materials engineering for silicon technology. Silicon is still the principal material for both digital and analog ICs. To have acceptable yields during manufacture, the number of defects, either present in the starting material or introduced during processing, must not exceed one per wafer per masking step. Such a stringent requirement has generated numerous discoveries and inventions: improvements in the methods of growth of bulk silicon material and epitaxial layers, development of lithographic patterning with feature-sizes below 1 micron and feature-size control below 0.08 microns, as well as improvements in plasma processing, etching, metal bonding and encapsulation technologies and techniques.

6.2.2.2 Economics

Consumer and communications electronics is expected to grow from \$ 745 billion today to about \$ 165 trillion by 2003. The mainstream technologies in semiconductor manufacturing are approaching a turning point caused by the increasing gap between the growing complexity of the ICs and manufacturing processes, on one hand, and the acceptable cost, on the other. For this reason, in the future the number of devices per chip may be constrained not only by the capabilities of the technology, but also by the

economics. The cost of a typical fabrication line of the future for making complex chips might be as high as \$ 10 billion. Therefore there may be only a few, say, five to ten, such lines in the entire world.

A semiconductor manufacturing factory of the future will be characterized by a high level of automation, simplified manufacturing processes, a high level of flexibility, high yields, clean chemistry, high-quality (particularly, contamination free) manufacturing, continuously improving capital productivity and appreciation of modularly upgradable equipment designed for very high reliability.

As the actual (and required) yield increases, production testing (which even today weighs 30% ~ 35% of the IC total production cost) will increase considerably. An emerging technique is concurrent testing. This technique will very likely become an industrial reality in several years.

Despite the growth in chip complexity, the design time for complex and highly packed chips will continue to decrease. This will be due, first of all, to the growth in the capabilities of computer aided design(CAD) equipment and software. By the year 2010, the design of a complex chip will, most likely, take just a few days and will be relatively inexpensive.

6.2.2.3 New frontiers

Progress in microelectronics results in greater computer power and larger storage capacity at lower cost. Long-term projections indicate that the microelectronics scene will continue to be dominated by complementary CMOS technology well beyond the year 2010. Over the last years, complementary CMOS technology has steadily taken over applications that were the

domain of bipolar devices. This trend is expected to continue as CMOS and bipolar CMOS (BICMOS) technologies penetrate the intermediate systems. Bipolar emitter-coupled logic (ECL) will continue to dominate high performance and supercomputer system applications, although gallium arsenide (GaAs) is beginning to find some applications in the supercomputer field.

The use of GaAs might be limited, however, by the availability of the appropriate software to operate superfast systems, and the advantages of GaAs semiconductors might be largely lost because of packaging and interconnection delays. Even if the software problem is solved, the high cost will, probably, confine the use of GaAs to very high performance systems. It is an obvious challenge for a materials and packaging engineer to reduce the delays to such an extent that the application of GaAs becomes technically and economically feasible. In addition, it will be the responsibility of a reliability engineer to solve numerous problems associated with thermal stresses in, and thermal management of GaAs-based chips.

The emergence of stacked 3D silicon might have an impact on future microelectronics systems integration. Volume manufacturing of this technology is now in place. The performance advantages of stacked 3D silicon are very large: the ultra high scale density results in factors of hundreds to thousands in both speed and power, when ICs are designed for 3D applications. Stacked 3D silicon has already had a major impact on microelectronics systems and products, into which it has been integrated. Examples include solid state data recorders, digital signal processors, massively parallel processors, artificial neural networks, image processing and imaging sensors. The present status and projections show

that, as volumes rise, no significant premium will be required to incorporate stacked 3D silicon into standard products. With the coming next generation of 3D stacked silicon, 10 to 1 000 layers of ultra-thin, low power circuits with thousands of inter-layer interconnect will comprise an entire system in a single cube. Power reduction will be a key issue for many applications. Concurrent development of adequate hardware and software will be necessary.

6.2.2.4 Nanoelectronics

The smaller the devices are, the closer we come to entering the world of quantum mechanics, where the rules of physics differ drastically from those in the macroscopic world. A possible new frontier in IC technology, based on the laws of quantum mechanics, is the application of devices based on ballistic transistors, and single-electron devices.

In the past several years, it has been established that solid state systems can be manufactured in such a way that the electrons travel freely (“ballistically”) for appreciable distances (about several microns) before they collide with crystal impurities. This provides a possibility of making a new type of transistors — “ballistic” transistors. It is possible that “ballistic” circuits could operate more quickly and dissipate less heat than conventional transistors. If this direction is successful, the new devices might begin to gradually displace the conventional ICs.

As to single-electron devices, these will be based on semiconductor quantum dots (also known as nanocrystals or nanoclusters). In a single-electron device, the data will be transmitted not by tiny currents, but by single electrons that hop from one quantum dot to the next. The wave mechanics of

electrons (the laws of quantum mechanics) will be applied to analyze the behavior of single electrons. Physicists in France have recently detected a phenomenon of decoherence by contriving systems that can be in two quantum states at the same time, albeit briefly. The use of this phenomenon is thought to enable one to provide transition from the quantum world into the macroscopic world. It has been found also that quantum dots, similar to bulk semiconductors, can radiate light when galvanized with protons or with an electrical current. Therefore semiconductor quantum dots exhibit potential use in future optoelectronic devices.

Back in 1985, Japanese scientists (Yokoyama et al.) proposed and demonstrated a resonant-tunneling hot electron transistor (RHET). In 1993 they developed a multiemitter RHET. This development enables one to make a SRAM cell and multi-input logic gates using a single transistor. Although the physics underlying quantum well transistors is not yet completely understood, the development of quantum tunneling transistors allowed the Japanese physicists to create a technology needed to fabricate precisely controlled quantum well structures. This technology has been successfully used to develop heterostructure transistors and quantum well laser diodes. RHETs cannot yet be used in practical applications. One reason is that the RHET should be cooled down to about 77 K. and its functionality is insufficient to replace conventional semiconductor devices. To make RHETs practical for future use in cryoelectronic systems, integration with complementary HEMT logic circuits or development of new architecture circuits will be necessary. Other important areas of research in this direction are development of room temperature quantum functional bipolar transistors and development of quantum

box devices based on RHET technology. These research direction will also be useful in the development of photonic devices, such as quantum dot lasers and new photonic memory devices.

6.2.2.5 Heterostructures

The advent of semiconductor heterostructures (such as, say, III/V and Si/SiGe) has opened many opportunities for novel microelectronics. It is anticipated that heterostructures will sooner or later find their application in the micro and optoelectronic market. Heterodevices are expected to be used as heterobipolar transistors, heterofield effect transistors, resonant tunnel diodes and, to a lesser extent, as some optoelectronic devices. Considered figures-of-merit are frequencies, transconductance, noise at high and low frequencies, threshold voltage, power delay, threshold current and quantum efficiencies. In the case of logic, the trend goes to nanoscaled devices and ICs targeting nanoelectronics. Heterostructure layers allow vertical nanoscaling and thus give an additional degree of freedom for design and optimization. It is viewed as an attractive challenge to solve some related technological problems like thin, low thermal budget oxides, defect free buffer layers, etc. Si/SiGe is now on an upswing in electronics and photonics.

The current developments in the physics of heterostructures focus on the following directions: physics of the heterostructures materials and systems (fabrication processes, materials issues, bandgap discontinuities, approximations and their breakdown at small dimensions); the follow-on of existing physics topics now carried out in heterostructures; the novel physical concepts and experiments that might be carried out in the future in heterostructures (a bottom-up approach to new physics driven by

the capacity of heterostructures to act as microlaboratories for demonstrating new physical concepts, even originating from other fields of physics); the novel physics to be done in heterostructures in order to answer new demands, be it on general motivations (i. e. advances towards the limits of computing or communications) or on evaluating the physical limits of existing heterostructure concepts (i. e. defining the physical limits of electrical to optical energy conversion in potentially thresholdless lasers or in LEDs).

6.2.3 Photonics

6.2.3.1 Why light beams?

Photons can do many things better than electrons. Light beams have practically unlimited information capacity (very broad bandwidth), very low transmission losses, do not dissipate heat, and are immune to cross-talk and electromagnetic interference. Therefore it is very likely that in the next century microelectronics will be largely replaced by optoelectronics and photonics. Fiber optics, for instance, has already taken over many functions of electronic systems. The world production of optical fibers has grown from 3.5 million kilometers in 1988 to 44.0 million kilometers in 1997.

6.2.3.2 Applications of photonics

Although lightwave communications technology has been used commercially for more than 20 years, it is only in the 1990s that cost reductions, induced by performance requirements and technology, have caused a dramatic increase in the use of photonics. Application of photonics engineering today can be categorized into two major areas: information (which includes communicating, sensing and displaying information, as well as

computing, i. e. processing and storing information) and materials processing (cutting; joining; machining, including surgery; patterning; etc.). Photonics-based devices have some use also for power generation (e. g. solar cells).

The world of telecommunications is rapidly moving from copper wire networks to fiber optics, and optical disks have become the major means for storing information. Photonics dominates displays and cameras, and is extensively used in printing. There is significant use of high-power lasers for materials processing. Optical lithography is a key tool in the electronics industry. There is growing use of photonics in backbone distribution networks, such as cable TV and personal communications networks. There are major potential applications of photonics in local networks, especially as broadband and multimedia services grow. It is expected that optical fibers will be widely used in the local telephone networks. It is expected that in several years from now lightwave communications products will be effectively used for in-the-home and automobile communications.

6.2.3.3 Photonics market

The photonics market is currently estimated as about \$ 60 billion per year. Displays (including laser printers) and memory disks (mostly compact and other laser disks) each account for about one third of this market. The third largest segment is due to telecommunications (about \$ 4 billion). The overall world photonics market is predicted to be in the \$ 100 ~ 200 billion range in 2005. Trend data indicate that we not only have at least another 20 years of progress, but also that today's most advanced research photonics systems utilize only about 0.1 % of the physical limits of lightwave technology.

Photonics is still an immature technology and suffers from a number of aspects, such as high cost, which currently inhibits the introduction of photonics in many markets. If the cost issues are solved (e. g. packaging of optoelectronic devices with fibers), photonics has the potential to enter a very broad range of new markets. At the same time, being an immature technology, photonics has the potential for significant breakthroughs, both in technology and in new areas of application. There is a crucial contest underway for franchises in providing broadband fiber optics communications (voice, data, video) to homes and offices. The markets for different photonics products and services are converging — and in such a convergence lies both uncertainty and opportunity.

6.2.3.4 Key photonics technologies

The key photonics technologies (features) that have fueled progress in photonics are: (1) creation of ultra pure glass fibers that attenuate optical signals by no more than 0.2 dB/km; (2) invention and development of reliable diode lasers; and (3) invention of erbium-doped optical amplifiers (EDOAs).

At the system level, a key commercial technology, the Wavelength Division Multiplexing (WDM) technology is underway. This technology is based on the realization of the fact that contemporary lightwave systems utilize only a small fraction of the potential bandwidth offered by optical fibers. Using WDM, multiple wavelengths (“colors”) of light, each carrying an independent stream of information, are concurrently sent down the same fiber. These “colors” do not interfere with one another, so that four, eight, 16 or more channels can use simultaneously the same fiber waveguide. Once multiplexing, di-multiplexing and

precision wavelength control are achieved, cable bandwidth is instantly multiplied by the number of channels used. The major current photonics technologies are: optical fibers; semiconductor optoelectronics devices (light sources and detectors); optical storage media; cathode ray tubes and liquid crystals (for displays); and high-power lasers for materials processing.

The key future photonics technical developments with the potential to have a significant impact on many fields of "high-tech" engineering include : improved manufacturability and yield for low cost, effective packaging and integration for optical subsystems; development of novel optics for advanced photonics systems; improved high-power light sources for memory, displays and sensing; improved lasers and detectors for networks and sensing; and development of new optical materials (e. g. for storage information).

The current effort in photonics engineering is in the following major fields: photonic local area networks; optical interconnects; lithium niobate and semiconductor switch arrays; silicon optical bench technology and passive components; improved laser technologies; optoelectronic ICs; and optoelectronics and photonics packaging.

Photonic devices rely on a compound semiconductor technology that, to date, has been most successful using the III-V semiconducting materials, such as GaAs and InP. A large family of them, direct bandgap materials, are ideally suited for light generation: laser and light emitting diodes (LEDs) and light detectors. Molecular beam epitaxy technology has become the workhorse for the mass production of lasers, such as those used in compact disk (CD) players. This technology can grow thin films

as thin as a single molecular level, uniform across the entire wafer, and reproducible from wafer to wafer.

6.2.3.5 Fiber optics communications

In the simplest fiber optics communication link the basic electrical signal is converted into an optical signal in a lightwave transmitter. At the destination, a lightwave receiver provides a digital data stream for further processing or other use. The information is transmitted by encoding it in the "ones" and "zeros" of digital data, with the light source turned on or off to represent "one" or "zero" ("yes" or "no"), respectively, i.e. similar to the function of the p - n junction in a semiconductor device.

The efficiency (capability) of lightwave transmission is measured by the product of the bit rate and the distance that the optical signal can travel before needing regeneration (amplification). Today's in-service systems can transmit information at the rate of 3.4 billion bits per second over a single fiber for distances of many kilometers. The lightwave transmission capability has been doubling every year for well over a decade. It is expected to continue to double every year for the next two decades. The break-even distance, at which the use of fiber optics becomes economically advantageous over copper wires has been increasing by an order of magnitude every several years.

Lightwave systems running at bit rates up to 1.7 gigabits per second are deployed in metropolitan areas, as well as between large cities, to connect central offices and major switching nodes. Lower bit systems (below 90 megabits per second) are operated as subscriber loop carriers linking central offices to remote terminals in business and residential areas. Single mode fibers with low transmission losses (below 0.2 dB/km) are now commercially

available.

Fiber to the end user—to the home and the business — will be one of the last steps in fully achieving an all-fiber communications network. This step is being stimulated by the wish to provide video services to end users, and by the introduction of the new digital formation broadband Integrated Services Digital Network (ISDN). Once this system (described in greater detail in Section 4.4) gains momentum, growth of the residential fiber network is expected to be very rapid. Within the next 15 years or so, the fiber network could grow to 100 million access lines.

A significant impact to international telecommunication services has been given by large capacity and highly reliable optical fiber submarine cables. Around the world, demand for the digital telecommunication service is growing and more than 250 000 km of optical fiber submarine cable have been laid. Fiber optic submarine cable systems are viewed as an opportunity for emerging markets to access the global highways and super highways. The construction of submarine fiber optic cable systems as a transmission medium has enhanced the quality and variety of global telecommunications services while increasing competition, reducing prices and perhaps, most significantly, stimulating overall economic growth of different geographical regions. The first undersea optical cable system was installed for service across the Atlantic Ocean about 15 years ago. In the Pacific and Asia, many optical fiber submarine cable systems are now being constructed. The advanced and large capacity cable system, TPC-5 cable network, adopting 5 Gbps optical amplifier technology, was laid in 1995/1996.

Bell Labs has recently demonstrated, as a new frontier in fiber optic telecommunication, a technique for transmitting data over fiber optic lines at a rate of one terabit per second. At that speed, which is more than 400 times faster than the current technology, the text of a daily newspaper could be transmitted in a mere second. The new technology uses a combination of multiplexing techniques, in which polarized light beams of slightly different wavelengths are used to carry data simultaneously over the same glass fibers. Although the basic physics of the technology has been proven, several substantial obstacles exist. The first is that the light signals are weak—in multiplexing, a beam must be split into, say, 100 different signals. These signals need to be reamplified at intermediate points, for example, every 50 miles. But that game can only be played for so long. After several such amplification processes, the signal-to-noise ratio becomes unacceptably small, and the signals cannot be received error-free. Therefore, unless further advances are made, a terabit per second system will not be suitable for long distances, that is, for trans-oceanic transmission. Furthermore, the multiplexing techniques need further refinement. The fibers themselves are not the problem because they inherently have very large bandwidths. The problem is in getting all the bits on and off the fiber when you are trying to transmit terabits of information per second. If the technological kinks can be worked out, terabit per second fiber optic systems could accelerate the era of network computing, making trivial the transmission and reception of large applications and huge quantities of data.

For future fiber optics telecommunications, new materials with low optical losses, well below 0.20 dB/km, should be

developed. As to the mechanical performance of these materials, they should possess sufficiently high long-term static fatigue strength and good protection against hydrogen and water. An improved long-term reliability of silica material in optical fibers can be achieved by using amorphous carbon films or metallizations, which hermetically seal the fiber, thereby increasing its long-term reliability. There are, however, many other mechanical performance problems associated with the design, manufacturing and use of "hermetic" fibers. Examples are: understanding and minimizing, if necessary, the thermally induced stresses in metallized fibers; development of design guidelines for metallized fibers subjected to lateral and/or angular ends misalignment, design of partially metallized fibers considering the stress concentration at the end of the metallized zone, etc.

6.2.3.6 Lasers

The progress in fiber optics communications is due not only to the development of low loss optical materials, but, even to a greater extent, to the invention of the laser, which is used as a light source in a fiber optics communications system. State of the art high performance semiconductor lasers that convert electronic signals into light pulses can now produce well-defined light pulses that have a duration of less than a nanosecond. Improved lasers that emit light of "rock-steady" wavelengths will usher in the age of coherent optical transmission, which sends concurrently different data on different wavelengths ("colors") through the same single fiber (WDM technology).

The invention of the laser is one of the outstanding technical achievements of the second half of the twentieth century. Although the advantages of lasers for communications were recognized

immediately after the invention of the laser, practical applications awaited the development of a host of supporting technologies. Paramount among these was the development of fiber optics suitable for long distance transmission, and of reliable, low-cost diode laser devices. In telecommunication systems, semiconductor lasers are being used as light sources and detectors. Today, fiber optic communications systems using laser technologies encircle the earth.

Vertical-cavity surface-emitting lasers (VCSELs) are promising devices for achieving effective optical interconnections. In recent years, this technology has made rapid progress. VCSEL array link modules with increasing speed and high performance VCSELs, such as low-threshold or high wall-plug efficiency ones, have been developed by American consortia (OETC and POLO).

Lasers are also used in various information technologies, including generation, transmission, storage, reproduction and display of information. Information technologies account for about 90% of all diode laser sales worldwide, as well as for a significant fraction of non-diode sales. Current developments could lead to further expansion of the use of lasers in information technologies. Examples are cross-talk free semiconductor optical amplifiers, 3D solid state displays and increased integration of electrooptical devices. These devices will continue to play an important role in information technology in the next century.

High-power laser diodes are currently used as pump light sources and for materials processing. Higher power, up to 5W and improved beam-shaping techniques are creating new opportunities for high-brightness diode-laser devices in communications, industry and medicine.

KrF excimer laser sources are being introduced into leading edge manufacturing equipment for integrated circuit lithography at 0.25 ~ 0.30 microns feature sizes. The most likely path for advanced IC lithography over the coming decade will involve excimer laser sources and projection optical systems extending all the way down to about 0.13 microns feature size. Such technology will be used to manufacture 4 Gbit memory chips, microprocessors and other integrated circuit systems with upwards of one billion transistors on a chip. Still, optical lithography is facing difficulty in resolution for the isolated pattern arrangement under a quarter micron region, even if KrF excimer laser lithography is used. It is expected that in the foreseeable future the electron beam lithography will play a complementary role for optical lithography. The mix-and-match strategy will open the door for large volume production use of electron-beam lithography.

6.2.3.7 Polymer lightguides

Japanese scientists (Koike et al.) have recently introduced a large-core, high-bandwidth and low-loss graded-index polymer optical fiber (POF). Using a new material, perfluorinated polymer, the attenuation of polymer lightguides has been lowered to about 40 dB/km. This enables up to 1 000 m transmission. A large core (0.2 ~ 1.0 mm in dia.) in a polymer lightguide enables one to use inexpensive polymer connectors. These can be prepared by injection molding, thereby eliminating the problem of modal noise. On the other hand, the development of the perfluorinated amorphous polymer base opened the way for the high-speed polymer optical fiber network. However, thermal stability and long-term reliability are two important issues for the viability of such a network. Since the refractive index profile

formed in the fiber is due to the distribution of the dopant molecules, the migration of the dopant has to be inhibited even at a high temperature atmosphere in order to maintain the refractive index profile. Long-term stability of the fiber strongly depends on the dopant used. It has been shown that the refractive index profile can indeed be maintained through high temperature aging by selecting the appropriate dopant.

Japanese scientists have proposed also an optoelectronic multichip module (OE-MCM), whose functioning is based on using fluorinated polyimide optical waveguides. The module can be manufactured on an optoelectrical substrate (OE-substrate), incorporating photodiode-to-waveguide coupling, and fiber-to-waveguide coupling. The OE-substrate is composed of fluorinated polyimide optical waveguides fabricated onto a copper-polyimide multilayer electrical substrate. Waveguide-to-photodiode coupling is achieved by deflecting the propagating light with a total internal reflection (TIR) mirror, and self-alignment of the photodiode is achieved using micro-solder bumps. The TIR mirror showed low reflection loss of less than 1.5 dB.

6.2.3.8 Optical amplifiers

Although optical fibers are used because of their low transmission losses, regeneration (amplification) is still necessary to compensate for transmission and splitted losses. If a "conventional" (electronic) amplifier is used, regeneration of the optical signal is needed every 80 ~ 120 km. In such an amplifier, a photodetector converts the attenuated optical signal into an electric signal, and high-speed electronic circuitry then reshapes and amplifies the electrical signals, which are reconverted back to light. Then a laser further transmits the regenerated optical signal.

The erbium-doped optical amplifier (EDOA), invented in 1987, is the major photonics breakthrough of the last two decades. This amplifier, which revolutionized optical fiber communications, controls light by using light, i.e. amplifies optical signals without converting them to electronic ones with subsequent reconversion of the electronic signals to light. In the predominant type of EDOA in use today, a 980 nm laser source ("pump") amplifies the signal in an erbium-doped fiber at 1 550 nm. EDOAs are currently used (1) in repeaters to boost optical signals periodically in longdistance systems (such as submarine ones), (2) to boost the power of the transmitter, (3) to increase the span length in transmission systems, (4) to compensate for splitting losses in networking systems, and (5) to enhance the sensitivity of the receiver. Just one EDOA can boost signals carried by many "colors" of light. This is especially important if capacity-building WDM technology is employed. EDOAs are substantially more efficient than electronic amplifiers and enable one to transmit information over very long distances without new regeneration. When used in trans-oceanic (undersea) cable systems, EDOAs are expected to enable signals to cross the oceans without regeneration.

6.2.3.9 Optical interconnects

Emerging optical interconnection technologies offer both new capabilities and packaging challenges for the systems designer. These emerging technologies include parallel optics, high bandwidth plastic optical fibers, and new VCSELs, which cover an increasing range of wavelengths. Parallel optics offers low-cost interconnection with perhaps the best use of backplane space. Plastic optical fibers (POF) have been demonstrated with high bandwidths above 1 GHz km and low loss between 850 and 1 300

nm. VCSELs will soon be available over a range of wavelengths spanning the spectrum from at least 650 nm to at least 1300 nm. Never before have systems designers of computing and telecommunication systems had so many choices for optical interconnection and different packaging approaches. Goals for packaging may include the ability to make very small systems, the ability to distribute systems to solve thermal problems, achieving good fiber management, getting the most bandwidth on and off and printed wiring board occupying the least space and disrupting backplane routine area the least, being able to insert boards having both optical and electrical interconnects without manual intervention with connectors, and the ability to flexibly manage heat dissipation in a system and on circuit boards.

The use of optical interconnects between processors, boards, chips, and even gates (devices) can drastically increase interconnection speeds. Optical interconnects will be vital to the development of high performance digital systems. The large bandwidth of optical fibers provides much higher speed and larger transmission distances than conventional electrical interconnects. Optical interconnects can also provide high parallelism because they are immune to cross-talk. Highly parallel board-to-board and chip-to-chip interconnects using 2D array devices and free space optical coupling are particularly appropriate for dealing with the massive amounts of data encountered in images (displays).

Lasers have been used in experimental optical interconnects on chips and from one circuit board to another. With the boards spaced about an inch apart, for example, light could travel through free space from lasers on one board to detectors on another to carry information between the boards. In the high-

speed large capacity digital systems of the future, optical signals going in and from integrated optical functional devices should permit logic processing unlimited by transmission. Combining these devices with free space parallel interconnects will provide powerful functions for processing extremely large amounts of data. This is expected to lead to the development of optical coputers and photonic switching systems.

The development of optical interconnect devices will proceed through a synergetic collaboration among material and processing technologies, design and fabrication of integrated optoelectronics, and optoelectronic packaging technology.

6.2.3.10 Nonlinear optics

A major objective of the research on nonlinear photonic materials is to produce all-optical switching and signal processing devices that will operate at speeds much faster than can be attained with electronics. However, to gain acceptance, high-speed photonic devices must satisfy several stringent requirements, some of which are low linear and nonlinear loss, and low operating power. Until recently, silica was the only "nonlinear" material used successfully in high-speed signal processing at practical power levels. However, fiber devices have serious drawbacks for some important network applications. At the present time, the most promising and exciting alternative to fiber devices seems to be semiconductor optical amplifiers. It has been discovered recently that active semiconductor devices can be made to perform photonic signal processing functions very effectively at low input power and at speeds not exceeding 10 ps, i.e. much faster than the carrier-recombination time.

6.2.3.11 Computing with photonics

Some consider photonics to be the key that will lead to lightning fast optical computers using light instead of electrons to perform calculations. Recent advances in optical data links point the way towards broader application of photonics in the computers of the future in which photons are expected to be used both for storing and processing information. It is very likely that computer input signals will be from optical fibers, thereby contributing to the development of higher performance computers. Encouraged by the successes in computer-to-computer interconnection and in computer peripherals, computer scientists are beginning to explore ways to use photonics within the computer itself. New concepts in computer architecture take advantage of optics and optoelectronic to provide interconnection paths in the dimension perpendicular to the computer circuits. This requires devices such as 2D arrays of light modulators, surface-emitting laser diodes, detectors and beam-steering holograms.

Although new devices for optical computing range from those operating at longer wavelengths down to the visible wavelengths, the major effort has focused on the very-near-infrared range (0.8 ~ 0.9 microns). This is due to (1) the availability of GaAs laser diode technology; (2) the availability of silicon detectors (which are cheaper than longerwavelength detectors and can be more easily integrated with silicon electronics), and (3) the realization of a smaller diffraction-limited spot size than is possible with the longer-wavelength devices.

The availability of the technology to produce and integrate computer components can make photonic computers a reality in the near future. Microcomputer systems are doubling their processing power every year. It is expected that computing 20

years from now will still depend on the microcomputer, most likely based on photonics, and on the distributed processing power of networks ("clusters") of microcomputers with high processing power.

6.2.3.12 Holographic memory

Holographic memory storage is a promising technology for digital data storage since the late 1960s. The inherently 3D nature of holography makes high-density storage possible. Many "pages" of data can be superimposed in the same volume of material. In addition, holography has a potential for fast transfer rates and, in some cases, for random access times. Fast transfer rates are possible because the data are stored and recovered in parallel (typically, one million bits at a time). On the other hand, some storage architectures allow for non-mechanical access of data. This can provide sub-millisecond random access times.

IBM has recently commenced a \$32 million holographic data storage project. The technology calls for data to be stored holographically as "pages" of bits in an optical medium such as a crystal. Early results from experiments portend that holographic technology could be used to store 12 times as much data at the same cost as magnetic disk storage. Furthermore because a laser is used for writing and reading the data, holographic storage promises input/output rates 10 times faster than its magnetic counterpart.

Bell Laboratories has just made significant strides towards establishing the commercial feasibility of holographic data storage, using a photopolymer medium as a new type of storage material. In the Bell Labs design, the data, during recording, are encoded with error correction and channel codes, and presented to the

optical system as pages of binary data using a device called a spatial light modulator (SLM). The SLM consists of approximately one million pixels, with each pixel representing a binary 1 or 0 by effectively passing or blocking the light. This modulated beam is then coherently interfered with a reference beam that can use light emitting diodes. The interference pattern of the reference and the modulated beams is then stored as index perturbations (the hologram) throughout the entire volume of the medium. To achieve high densities, more than one hologram is stored in essentially the same volume by changing some characteristic of the reference beam. This process of superimposing holograms in the same volume is called multiplexing. The traditional multiplexing technique is to change the angle of incidence of the reference beam so that each hologram is stored with a unique angle. The separation between holograms relies on the coherent nature of the hologram to reconstruct in phase throughout the volume only if this angle is correct. Since the required separation between holograms decreases with thickness, more holograms (larger density) can be stored in thicker materials. To read out the stored data, a reference beam with characteristics matching those used during recording (color, angle of incidence and position relative to media) illuminates the media and diffracts off the stored index perturbations to reconstruct the stored modulated beam. The bits are then detected in parallel by a multi-element detector. Since all the one million bits are detected and stored in parallel, the transfer rates can be as high as 1 Gbit per second. The recovered "page" is then processed using the channel and error correction codes to reconstruct the original data. A density of about 50 bits per micron squared has been achieved in a photopolymer medium. Materials

in hand will result in 5 1/4 in. disks with 125 GB of user capacity, more than 30 MB/sec read rates, random access, and low-cost removable media.

6.2.3.13 Laser beam welding

Pulsed laser welding has gained wide acceptance in photonics packaging as a highly reliable material joining technique.

Laser beam microwelding (LBM) represents an alternative joining technique to soldering or, sometimes, to adhesive bonding. The main advantages of LBM technology are high temperature strength, reduced manufacturing time and simplified material separation at the end of the life cycle. The main drawback is difficulty in controlling sensitive LBM process behavior.

Pulsed laser welding is attractive for applications that demand minimal size of the heat-affected zone and accurate positioning of the weld area. Such requirements are necessary, for instance, to ensure good coupling between a laser diode and an optical fiber. The technique has found wide application in laser packages for single mode transmission. In such packages, sub-micron motions of the package components can cause significant coupling losses. In particular, if the laser moves laterally by 0.5 microns, there is a loss of 1 dB, and even a 0.1 micron movement can result in 1 % loss.

6.2.4 Other Future Technologies

6.2.4.1 Video services and visual communications

Our desire to be entertained will become a powerful marketplace incentive for the evolution of an all-fiber network, including fiber-to-the-end-user. Several trends suggest that

entertainment video will become both interactive and available on demand. For example, viewers will be able to select movies from a list of box-office hits, and order these selections in any quantity and sequence at any time. Some movies and live programs will be designed to be interactive with the viewer.

It is anticipated that by 2010, the network of interactive telecommunications and entertainment will be widespread, and will enable users to indirectly and remotely experience a place or an event in all dimensions. This network will allow people to participate in meetings without travel. Two-way telecommunications connections will be used, in whole or in part, to provide video communication among friends and business colleagues, as well as to provide a wide variety of educational and entertainment video material to homes, schools and offices. Consumers will originate and create video material, as well as interactively select and receive it. It will become possible to attend meetings and other events at remote locations by simply flipping channels or to travel without leaving home. Owing to 3D imaging, a person will be able to turn his head and see different aspects of a particular scene.

The visual communications system of the future will be an array of technologies, components, communications networks, products and services. These will meet a broad variety of customer needs, spanning the gamut from emotional closeness with family members to interaction with customers, suppliers and co-workers. These services will also be helpful in teaching the latest concepts and technologies in various fields of knowledge in a convenient and efficient manner. Special view and multimedia chips will give equipment manufacturers the core technologies needed to bring

affordable, full motion video and CD-quality audio to desktop microcomputers and workstations.

6.2.4.2 Displays

The global market for flat panel displays (FPDs) is currently worth about \$ 10 billion, with increasing demand forecast for the future. The current dominant technology is liquid crystal displays (LCDs). LCDs account for about 90% of the high information content display market. It is a nonemissive light valve technology requiring backlighting and having tremendous advantages in terms of power consumption and cost. However, LCDs still have intrinsic problems associated with production and operation. The major disadvantages of LCDs are viewing angle and temperature limitations. The competing emissive technologies are electroluminescent, plasma, light emitting diode/diode laser displays and, especially, field emission displays (FEDs).

FEDs are the most competitive second-generation technology. They have several significant advantages over LCDs: superior performance in terms of low power consumption, high brightness, sunlight readability, full color, high contrast, fast response time, large viewing angles, and change in contrast or color as a function of angle. FEDs also possess a large dynamic range of light output including good dimmability, fast refresh, very high spatial resolution, and high pixel count in a windowpane thin package that has a viewing area that is from very small (e. g. camera eyepiece) to very large (60 inch diagonal). FEDs have the potential to be scaled up to large screen sizes with low cost and high yield. FEDs are being developed for a number of applications including military and commercial aircraft cockpit displays.

6.2.4.3 Speech recognition and processing

Speaker-independent automatic speech recognition and synthesis will become commonplace. Intelligent machines will be able to talk and listen much as people do. By the year 2010 we will be able to transfer personal speech characteristics across languages, to make possible customized speech translators. Speech in one language will be automatically translated in real time into a second language, and the translation might then be synthesized, also in real time, with the voice characteristics of the original speaker. It is estimated that by the turn of the century, speech recognition systems will understand more than 10,000 words, twice the number of words used daily by most people, with 95 % accuracy.

6.2.4.4 Network architectures

The trend towards distributed microcomputer-based processing power has led to ever increasing intelligence in telecommunications and information networks. Such network intelligence will make possible network resources on demand for increasingly widespread resources needed. For example, network users will be able to access the exact amount of bandwidth they need for specific applications, such as videoconferences, and get that bandwidth when they need it and for however long they need it. As part of such network resources on demand, most services will become pre-provisioned: the capability to provide a broad spectrum of services will be set up in advance. A user or service provider will activate the given service on demand.

Network architecture will accommodate a variety of multimedia applications using voice, data and images. The key to the efficient handling of such applications will be the mature and widely deployed digital formats of Integrated Services Digital

Network (ISDN), broadband ISDN (B-ISDN), and Synchronous Optical Network (SONET).

Narrowband ISDN will mature over the next two decades. This will change the way we work and play. The telephone service may be largely replaced by B-ISDN as the basic wired service. The major benefits of the B-ISDN are high-quality videophone and high-fidelity voice. It is expected that by the year 2010 B-ISDN will be widely available. This will make a wide range of possibilities feasible, including high-speed distributed data networking, networked supercomputing, as well as multimedia networking and video services. To meet the demands for making B-ISDN a familiar part of our daily life, high-speed data communication equipment will 100 Gb/s data throughput and Tb/s optical transmission systems will be required. This means that there will be a crucial need for high-performance and cost-effective packaging of optoelectronic devices.

Photonic transmission facilities will be based on the evolving international standard—the Synchronous Optical Network—(SONET), which defines standard network interfaces. Owing to SONET, telephone companies and end-users will be able to employ equipment from many different vendors without worrying about compatibility.

Existing telecommunication networks consist of optical transmission lines and electrical cross-connects and switches. As a result of the worldwide growth in traffic volume, there is a need to further reduce costs, requiring better utilization of fibers and high-capacity nodes. Both photonic technology and microelectronics will be required in future telecommunication networks and can be used in a complementary manner. Transmission bandwidth and

cross-connecting can be strongly enhanced by adding systems based on wavelength-division multiplexing to the existing systems (synchronous digital hierarchy and/or asynchronous transfer mode) that use electrical time division multiplexing. For the next few years, switching nodes will still remain in the electronic domain because of the immaturity of optical time-division multiplexing technology. Also specifically discussed are the realization, evolution aspects and standardization issues of optical networks/crossconnects. A demonstrator in which some functions are still realized by microelectronics can become reality within a short time. However, much more research and development is necessary to create an all-optical cross-connect product.

Photonic networks will be the basis of a future broadband Internet that permits easy access to a variety of multimedia services. Optical frequency and time division multiplexing techniques enable an optimum matching of the photonic network's capacity to the increasing communications demand, thereby reacting to the requirements of new services in a flexible way. Flexibility and reliability with high profitability are the crucial requirements and crucial characteristics of future photonic networks.

Cost-effective InP-based photonic integrated circuits (PICs) are expected to find wide application in future optical access networks. They can be used, particularly, in the following bi-directional schemes: Time Compression Multiplexing (TCM) at single wavelength; Wavelength Division Multiplexing (WDM) with two wavelengths for the upstream and downstream signals; and WDM transmission with two wavelengths in the fiber window.

There is also a strong need for wireless networking. Because

people are increasingly on the move, a large number of network end links will be wireless. Customers will have great mobility, and will depend on Personal Communications Networks (PCNs) to automatically direct calls to them wherever they are. A combination of wireless technologies and network-based messaging capabilities will make people reachable any time and anywhere, if they want to be contacted.

It should be pointed out that in many cases it is political objectives that stimulate competition today: privatization and liberalization are currently changing the world telecom and multimedia market. The continuous evolution in silicon processing and optical fiber technologies are enablers for next-generation systems and applications. Broadband Networks (ATM, SONET, SDH, etc.), Full Service Networks (ADSL, VDSL, HFC, etc.) and Wireless Personal Communication Systems (GSM, DECT, etc.) are only a few examples. It is envisioned that microelectronic system design will evolve together with silicon and fiber technologies, in order to keep system development manageable during the next decade.

The introduction of new types of interactive multimedia services (Internet, etc.) demands substantially increased capacity of the existing telecommunications network infrastructure. It has been proposed that polymeric microstructure technology be used for future low-cost, high-volume tele- and datacom components, especially in the fiber-to-home networks. This will bring down the cost of optical terminations, including optoelectronic modules, fiber optic connectors, splitters, etc.

6.2.4.5 Switching

A key capability behind the evolution of network architectures

is the evolution of network switching. During the 1990s we see the transition to broadband switching initially with software control for a stand-alone packet switch, which will provide broadband ISDN. initially at rates of 150 megabits per second. We will also see the transition to a B-ISDN switching module for a network switch. The switching evolution will continue to the year 2010 and beyond. It is expected that by the year 2010 switching systems will easily accommodate photonic transport, with fiber lines in and fiber lines out. It is expected that the actual switching fabric of the future switching systems will all be photonic.

6.2.4.6 Software

The major software problem today is to increase programmer productivity to meet exploding demands. Such an increase is expected to come mainly from the growing sense of previously developed and tested software modules. Standards for software and electronic systems should be pervasive and widely deployed in 2010. Software design will be, for the most part, automated in a way similar to today's computer aided design of ICs. Software jobs will be heavily concentrated on front-end work and eventually on new technologies like photonic-based software.

6.2.4.7 Wireless systems

The telecommunications industry is now in the midst of a second "wireless revolution" (after the famous inventions of Marconi at the turn of the century). The current revolution began with the invention of cellular communications technology. Cordless telephones are providing instant communications mobility in our homes and offices, and wireless local area networks (LANs) are being used for high-speed data networking inside buildings. Cellular services and paging systems provide

communications coverage virtually everywhere, and there are innumerable wireless remote control devices that make our lives easier and more fun.

One of every five new telephones in the United States is cellular. In the not too distant future people will find it quite ordinary to pay highway tolls by means of a wireless electronic toll-collection system, carry a notebook (or palm-sized personal digital assistant) with a wireless interface providing access to a range of personal information and messaging services, and to place and receive calls on portable personal telephones that access either an office wireless network, when inside an office building, or a public personal communications network (PCN) when outside the building. Emerging wireless systems and services will provide the technology to allow people and machines to communicate anytime, anywhere, using voice, data and messaging services through telecommunications.

The substance of the cellular concept is in dividing a service area into a number of relatively small "cells" and only using a subset of the radio channels in any given cell. These channels could be re-used in cells that were far enough apart, so as to keep co-channel interference below the allowable level. System capacity can be increased by reducing the size of the cells, thereby increasing the number of times a radio channel can be re-used. While the concept is simple, the implementation was both a regulatory and technical challenge.

The technical issues were solved in the United States by introducing the advanced mobile phone service (AMPS), which uses sophisticated signaling between mobile units and mobile switching centers to track and monitor mobile units, control the

radio transmission link, and perform real time hand-offs from one cell to another. With a mere 50 MHz a spectrum, there are now over 15 million subscribers to cellular services in the United States, and the number is rapidly growing.

Current ("first generation") cellular systems are based on analog FM radio technology. With the objective of improving quality and increasing capacity, second generation cellular systems will be based on digital radio technology and advanced networking principles, which will enable one to improve the quality and network capacity of a system. These "second generation" systems will increase the overall system capacity by a factor of three to 10 over existing analog cellular systems. The "second generation" systems use principles derived from ISDN and intelligent networks to provide end users with "seamless" roaming and access to a wider range of telecommunications services.

Maximizing spectral efficiency, while maintaining a high-quality communications channel under sometimes harsh and unpredictable conditions, could be the largest technical challenge the wireless industry faces. Unlike physical media, such as copper wire or glass fiber, which provide a somewhat predictable and controlled transmission environment, the free space environment for wireless systems is constantly changing and can result in co-channel interference (distortion by other nearby systems using the same frequency) or multipath interference (reflection of radio waves off buildings or other structures).

Wireless data systems are also receiving increased industry attention in connection with the advent of laptop and palmtop computer. Wide area, two-way packet radio and one-way paging networks are making it possible to send and receive messages and

access data wherever and whenever required. By allowing systems to be set up where traditional wiring cannot be used, wireless LANs introduce a new level of freedom for in-building data networking.

Currently, there is no wireless technology or system that will support all applications and end-user needs. Researchers and system architects are developing methods to support applications “seamlessly” across varying environments and technologies. Service zones can be used to categorize enduser environments, manage the appropriate wireless architectures, and plan the evolution of technology. Here are the four basic service zones: global/national service zone, mobile service zone, local/micro service zone, and indoor/pico service zone. International and regional standards bodis are considering how to develop standards that will allow for “seamless” service across all environments. These considerations range from a single, all-purpose radio interface, to a family of modular interfaces built from readily available components. With progress continuing in technology, standards establishment, and regulatory rule making, the vision of a “seamless” end-to-end global wireless service could become a reality within the next decade.

6.2.5 Conclusions

The key “ high-technologies ” of microelectronics, optoeletronics, photonics, computing and software will advance dramatically beyond today’ s powerful capabilities during the next decade. While microelectronics may possibly mature in about a decade or so (although mechanical and materials challenges associated with reliability and manufacturing of microelectronics

systems will remain for many years to come), the others have at least two more decades to go. Tomorrow's technologies will make us more productive, our lives easier and will give us more free time and more fun. These technologies will bring about three major things:

convenience—disappearance of wires; mobility; handheld computer power; and “smart” cards that let one to store, access, and use a wide variety of information

control—videophone monitor on or off, only calls from certain people get to you; reachable anywhere, or not at all, as you choose; and communications and computer devices that react to speaking, writing or typing

connections—widespread videophone capabilities for homes, offices, even pay phones; shop-at-home services that let you customize before you buy; real time language interpretation; interactive computer learning; and work-at-home option for millions.

Theoretical and experimental approaches, methods and techniques of mechanical, materials and reliability engineering are and will continue to be important and indispensable tools in the development of viable, reliable, convenient and, cost effective “high-technology” components, devices and systems of the future.

References

- 1 E-K. Aschmoneit. Quantum trap: nuclear physical effects in the components of the future, *Elektronik Praxis* 32(11)(1997): 88 ~ 90 (see also pp.92.94)
- 2 C Baack. Trends of photonic communications networks. *Frequenz* 50 1996(9 ~ 10): 192 ~ 197

- 3 D L Begley. Global trends and the future of laser communications. Proceedings of the SPIE—International Society of Optical Engineers (USA) 2990(1997):134 ~ 140
- 4 J C Bouley. InP-based photonic integrated circuits for future optical access networks. In: Proceedings of the Lasers and Electro-Optics Society Annual Meeting. Vol. 1. IEEE. Piscataway, NJ. USA. 96CH35895, 1996:286 ~ 287
- 5 J Carson. The emergence of stacked 3D silicon and its impact on microelectronics systems integration. In: Proceedings of the Eighth Annual IEEE International Conference on Innovative Systems in Silicon, Cat.No.96CH35996. 1996:1 ~ 8
- 6 P K Chatterjee, R R Doering. The future of microelectronics. IEEE Proceedings 86(1)(1998):176 ~ 183

6.3 Accelerated Verification of Digital Devices Using VHDL

As digital designs become more complex and increasingly include a processor on the chip, verification – and in particular the generation of testbenches – is becoming a bottleneck. This paper presents two aspects for improving the verification of microprocessors; program-less verification, and methods for handling large differences in abstraction level between a reference model and the actual design. Program-less verification is a type of pseudo random verification where the notion of a software program executing on the microprocessor has been abandoned. This removes some restrictions on the pseudo random data and instruction streams, and avoids limitations and bugs in software tools such as assemblers. A high level of abstraction is important when creating reference models, since it reduces the risk for introducing errors in the reference, as well as the effort for coding

it. However, the testbenches must then be able to cope with behavioural differences due to phenomena such as prefetching, speculative execution etc.

6.3.1 Introduction

With growing complexities of digital devices, the generation of a complete verification suite requires an ever increasing portion of the development time and resources. Today, even though extensive verification has been employed, devices with incorrect functional behaviour are occasionally manufactured and marketed. Typically, the erroneous logical implementation has not been detected due to inappropriate verification criteria. An important contributing factor to such oversights is that those preparing the test suite inadvertently tend to base it on anticipated device applications and therefore misinterpret the specification in the same way as the designers. An example of this could be when an embedded microprocessor is verified by cosimulation with the target application software. Another contributing reason is that the verification bottleneck lays in the time spent on developing the test suite rather than simulating it (Ref.1).

We have approached these problems by stimuli generation for microprocessors decoupled from the foreseen usage of the device. Our method relies on comparison of behaviour between different representations of the microprocessor. The notion of a valid software program is abandoned for the stimuli generation, effectively resulting in a program-less verification approach for microprocessors. Additionally, limitations associated with software developments tools are eliminated in the stimuli generation process. The ratio between simulation time and

testbench coding time is increased, allowing a larger number of errors to be found per time unit spent on stimuli development. These characteristics yield in an acceleration of the overall verification process.

For complex designs it is necessary to obtain a correct reference model that is used throughout the development (Ref.2). When describing a reference at a high abstraction level, the functionality is not obscured by implementation details which allows an easier understanding of the problem and reduces the possibility for introduction of errors. By applying the same stimuli to the test object as to the reference model, errors can be detected by comparing the obtained results. The testbench has to be able to adapt dynamically to the response from either model to tolerate correct but dissimilar behaviour resulting from the models being at different abstraction levels.

The stimuli is pseudo randomly generated and applied to the test object and the reference. Existing methods employ generators coupled with assemblers and other software development tools, producing pseudo random software programs to be executed as stimuli. Such stimuli have some disadvantages since they have to adhere to the limitations that are attributed to a valid and executable program. Instead of relying on existing pseudo random generators and software tools, we generate the stimuli during the actual simulation. A microprocessor is simply seen as a state machine being fed with instructions and data which are conceived on the fly, effectively resulting in a program-less verification approach. Any correlation between the specified function and its usage is thus avoided by completely abandoning the notion of a program.

To cope with different abstraction levels and to allow program-less verification, we have chosen to develop the complete testbench in VHDL. The testbench includes the reference model and the test object, and pseudo random generation of executed instruction and data sequences. Today, testbench development is more time consuming than the actual simulation. Methods increasing the simulation portion of the time spent on verification are therefore becoming attractive. The outlined approach offers an acceleration of the verification process by reducing the effort required for the definition and implementation of a stimuli for finding a given number of errors.

This paper starts with an overview of different functional verification approaches. It is followed by a presentation of our method and an in-depth description of how it was applied to a real project. Some relevant cases are then discussed. Finally, the conclusions from the experience are drawn.

6.3.2 Current Verification Approaches

6.3.2.1 Manual stimuli and inspection

In the past when designs were small, tests were conceived manually. For the stimuli, the simulator was commanded directly in front of the workstation, and the response in the form of waveforms was visually inspected on the screen.

Depending on the response, subsequent stimuli were conceived in an interactive manner. For example, if an error was detected, its effect on the verification could be reduced by avoiding to exercise that part of the design until the error had been corrected.

While this approach is manageable for small designs, its

drawbacks include:

- risk for confusing the actual behaviour as seen on the screen with what has been specified, especially when under pressure to complete the work;

- limited length of the verification since requiring the person to be present at the workstation while simulating;

- unsuitable for repeatable regression testing since interactive, and furthermore the quality of repeated simulations varies with the day-to-day performance of the person running it.

6.3.2.2 Extension for regression verification

As an extension to the above approach to handle regression testing, simulator commands were recorded in a file during the first simulation, and then played back for repeat simulations. Similarly, the response from the design was also stored in a file, for example as periodically sampled signal values. The manually inspected response from the first simulation was stored as a reference file, and files resulting from subsequent simulations were then compared to this reference using some sort of comparison utility.

This extension is suitable for verifying that the design behaves exactly as during the reference simulation, for example to verify that it has not changed due to a modification in the design. However, due to the rigid criteria to exactly correspond to the reference simulation, it is not well suited to handle minor changes in the behaviour that are still within the specification. An example of such a case could be a non-critical signal being active one clock cycle later than in the reference simulation. Even if this would be acceptable from a specification point of view, it would be signalled as an error by the verification, since it is not

sufficiently intelligent to know whether or not such a change represents an error. Since the stimuli is just a playback without any feedback from the design, changes could get the stimuli completely out of synchronisation with the actual state of the design.

6.3.2.3 Automated testbenches

With the introduction of hardware description languages, the designer was not only provided with a powerful tool for describing the design itself, but also with a tool highly suitable for verification. The concept of testbenches was introduced, comprising stimuli generators and checkers for functional correctness. The correctness checkers can be devised in ways allowing different, although correct, responses from different implementations or representations of a design under test. With checkers that abstract from implementation details and concentrate on functionality, verification at a higher abstraction level is made possible.

An example of how verification is performed at a higher abstraction level is when a checker comprises a bit-serial receiver that outputs the received data for further processing. A higher level checker can then analyse the received data without involving any implementation details of the communication protocol in the process. The checker is thus not bound by cycle or waveform-true comparison and can therefore accept low-level deviations. When cycle-true comparison is desired, as for regressive verification, the testbench can be complemented with signature generators and checkers based on multiple input serial registers.

In this approach the objective is to verify each specified function of the design in a systematic manner. However,

complicated interaction that might exist between different functions can be overlooked during the verification. The effort to find obscure errors is often a laborious task since the creation of explicit stimuli exercising all aspects of the design require careful analysis of the functionality and the implementation itself.

6.3.2.4 Pseudo random stimuli

To improve the chances to detect obscure errors in complex designs, pseudo random stimuli generation is often introduced to complement the systematic verification. Since it is nearly impossible to determine the correct response from a pseudo random stimuli, comparison with a correct reference model is often used (Ref.2). One important criterion is that the test object and the reference have been developed independently in isolation to eliminate unintentional correlation leading to systematic errors.

The same stimuli in the form of a program is executed on both the test object and the reference. The logical significance or meaning of the test object output, being address and data for a microprocessor, is arbitrary and is therefore not considered in the verification. The test object output is therefore only compared with the output of the reference. Any deviations encountered must consequently be examined, which normally requires that the behaviour of both objects is analysed since the cause of the problem could originate in the reference. As long as the behavioural differences are minute, the reference can be represented in a variety of styles, ranging from instruction simulators and behavioural models to real hardware.

For microprocessors, a common way of producing pseudo random stimuli is to compile a test program using a test generator, typically written in the C language (Ref.2, Ref.3). The generation

process normally requires the use of an assembler or other software developments tools, which introduces limitations on the produced test program. Since the test program is generated prior the actual execution of the instruction sequence, there are no possibilities for adapting the stimuli to the response obtained during the simulation. This limits the allowable behavioural differences between the test object and its reference. It effectively renders the method unsuitable for developments with large differences in abstraction level between the reference and the test object.

The usage of a software program as executable stimuli inadvertently introduces correlations between processors and their foreseen applications. Executable programs are bound to the following characteristics, limiting the efficiency of current pseudo random verification approaches:

- deterministic result and purpose of any subprogram;
- correlation between processor state and memory;
- an instruction fetch to an address expects a certain instruction to be returned;
- a data fetch to an address expects that previously written data are returned;
- exceptions and subroutines need to return to their origin;
- limitations in program flow, avoiding backward jumps resulting in infinite loops etc. (Ref.4);
- deterministic termination of the program.

Previous work has been performed to improve the efficiency of pseudo random verification. Instead of applying fully random data values to an arithmetic unit, weighting of the parameters can be performed to cover corner cases and improve the test coverage

(Ref. 5). To reduce the stimuli length, intelligent instruction selection can be performed taking into account the predicted internal state of the test object(Ref.6). To improve the quality of the generated stimuli, sequences can be generated that explicitly exercise different functional units simultaneously to increase the possibility to catch errors related to complex function interaction (Ref.4).

Little effort has however been made to remove the limitations inherent to software tools and the notion of software programs, an area that we will explore further in the following section. To achieve optimal verification, it is important that all the above approaches are being addressed.

6.3.3 Accelerated Verification Approach

We add a new aspect to the generation of the pseudo random stimuli than what has been previously addressed. Stimuli are generated on the fly without accepting the limitations of executable programs, also taking into account any deviations between test object and reference responses. All the necessary steps can be performed directly in VHDL without the need for external software development tools. It is basically a hardware designers approach to software-free microprocessor verification.

6.3.3.1 Program-less pseudo random stimuli

There is no resemblance of a software program in the testbench, only a sequence of instructions to be executed. The testbench does not implement a traditional memory in which a program is stored. When the processor reads instructions or data, it is provided a pseudo random value not related to the issued address. Two instruction fetches to the same address can thus

result in different instructions. A read operation will not return the same data as accessed by previous read or write operations. If an instruction was read from an address at one time, the test sequence might very well result in a data fetch from the same address later on, effectively eliminating the notion of data and program address space.

The deterministic flow of a program is abolished, allowing any sequences of jump and call instructions to be realised. If a jump is made to the same address as the current program counter value, the subsequent instruction fetch to the same address will generally result in the return of a different instruction. There is therefore no need for analysing jump sequences to prevent the infinite loops occur, since this risk is eliminated through the merge of the code generation and execution.

A subroutine call or exception handling routine does not have to return to the original program flow as is in real executable software, since there is no program flow whatsoever in our approach. As a result, the significance of data such as operational parameters is ignored, allowing e. g. a division by zero without disturbing the simulation.

6.3.3.2 Coping with different abstraction levels

As said in the introduction, for complex designs it is necessary to obtain a correct reference model that will be used throughout the development process. By describing the reference at a high abstraction level, the functionality is not obscured by implementation details which reduces the possibility for introduction of errors. However, the difference in abstraction levels will most likely result in behavioural differences between the two models.

For example, if the test object reads a group of operands in a different order than the reference, the computed result will still be the same. The same applies to prefetched instructions that are not executed as well as speculatively executed instructions which are discarded. An example of the former is when one of the models makes a prefetch of an instruction superseding a branch instruction which is not executed because the branch is taken, whereas the other model does not make the prefetch at all; the overall program flow still remains the same.

To cope with the behavioural differences between the two models, the testbench has to analyse actively in real-time the response from the test object and modify the stimuli as necessary. By applying certain tolerance levels for such deviations instead of immediately reporting an error and terminating the simulation, much of the associated analysis work can be automated.

6.3.3.3 Verification acceleration

The engineering task is then to develop a stimuli generating infrastructure that spawns the test suite and handles different behavioural responses from the two models. Arbitrarily long simulations can then be performed without human interventions. As previously mentioned, the simulation time is not the limiting factor, testbench development is the bottleneck. We move the test case generation to the actual simulation phase, allowing more test cases to be simulated with less development effort and therefore accelerating the verification process.

Although the method might seem to be simple, applying it to an actual real-life project is not always straightforward as will be shown in the following section.

6.3.4 Verification of a Mil-Std-1750 Microprocessor

The objective was to independent verify the correctness of the instruction set implementation of a microprocessor model, ignoring all signal properties such as timing and waveforms.

The object to be verified was a VHDL model, intended for board level simulation (Ref. 7), of an existing 16-bit microprocessor according to the MIL-STD-1750 instruction set standard. The model was described at a high abstraction level, although encompassing architectural bit-level details. The model was capable of executing approximately one thousand instructions per second when run on a SparcStation 10.

The board level model had already been extensively verified as part of the development. The foundry production test had been applied to the model, including an instruction test verifying the standard. In addition, specific tests for each instruction implemented had been devised, including testing of boundary conditions etc. The timing and bus cycle behaviour had also been verified. Finally, the model had successfully executed simplified application software in a board level system.

The reference model had been converted to VHDL from the design data base used during the developent of the existing hardware. The model was slow, only executing a few instructions per second when run on a SparcStation 10. Several iterations of the hardware had already been used for some years.

It should be noted that in this project the reference model was mush slower than the test object, since the task was to verify a fast board level model. This does however neither affect the approach nor the conclusions from this exercise, since in an active

device development the roles would simply be reversed for the two models.

6.3.4.1 The verification approach

The verification was performed without acquiring any implementation knowledge of the test object or the reference model. It was assumed to be sufficient to understand the instruction set architecture, treating the two models as pure black boxes. In this way, any correlation between the implementation and the verification was eliminated.

The testbench was implemented exclusively in VHDL, using no external software tools such as compilers, assemblers or linkers. Both models were simulated simultaneously on the same simulator. This was necessary in order to handle differences in timings, number of clock cycles per instruction, bus cycle behaviour, and in instruction and data flows.

We initially took a broadside approach to pseudo randomised stimuli generation, with no or only a few limitations during instruction and data flow generation. To allow long sequences of legal (and illegal) instructions the testbench had to be able to accept out of order instruction and data fetches, handle allowable mismatches between the two models, etc. The objective was not to halt the simulation immediately, but to try to recover and continue if the deviations upon detection of a discrepancy were found acceptable. The goal was to accelerate the verification process by reducing the time spent on the testbench development and instead spend the available time on simulation and error analysis.

6.3.4.2 Instruction set data base

To be able to handle behavioural deviations between the two

models, a detailed instruction set data base was described in a VHDL package. The data base was used during generation of the instruction and data sequences, as well as during error reporting that will be discussed later.

One of the primary functions of the data base was to allow the code generator to know what kind of information the microprocessor would expect during a bus access. For example, it was possible to identify whether to provide an opcode or a parameter to the microprocessor during an instruction fetch. This was used for several different purposes, one being the possibility to exclude arbitrary instructions from the verification if necessary.

The following list shows some of the information that was stored in the data base for each instruction:

- mnemonic, description and usage;
- opcode, number of operand accesses and data types;
- address mode and register usage;
- potential exceptions.

6.3.4.3 Instruction filtering

The instruction set data base allowed selection of subsets of the available instructions. It was possible to exclude an instruction from the test set if it was found working incorrectly. This allowed further verification runs without generating more errors of the same type. The verification could thus continue while the model was being corrected. The instruction in question could then be enabled again for further verification. This made it possible to target specific instruction types, e.g. jump, etc., or to select only instructions with previously erroneous implementations.

6.3.4.4 Instruction sequence generation

The generation of the instruction sequence was performed in

two steps. Firstly, a partial sequence was generated for the reference model which was executed an arbitrary number of instructions ahead of the test object. The generated instruction and data flow was stored together with associated address information in a first-in-first-out type of virtual buffer. The test object was then fed with the instructions and data from the buffer as they were accessed. The test object was not bound to execute the buffered instructions in the same order as they were stored. This was done to allow mismatches between the test object and reference behaviour which will be described further on.

The testbench decoded the bus cycles as generated by the reference model, to know what kind of information was expected. When a compound instruction was fetched, the whole instruction with all associated parameters was generated and remembered. As the reference fetched the remaining parts of the compound instruction, the already generated information was dispatched. The instruction set data base was instrumental in the generation of the reference instruction and data sequences. The selection of what instruction to generate was done pseudo randomly, but could be controlled by the aforementioned filtering capability.

The testbench also decoded the bus cycle accesses made by the test object. If an instruction was requested, the testbench would first see if the provided address was available in the buffered instruction sequence and would in such a case dispatch it. If the requested address was not first in the queue, the preceding instructions would be reported and removed from the buffer, allowing instructions to be dropped.

If the requested address was not in the buffer, a nonoperation instruction would be issued, for example to allow the test object

to have a different prefetch implementation than the reference. This also allowed the test object to generate an exception not previously done by the reference, without immediately upsetting the simulation.

A similar approach was followed for data fetches, but in this case buffered data were not immediately dropped to allow out of order operand fetches. Special provisions were implemented for atomic operations etc.

All the aforementioned mismatches between the two models were tolerated up to a user defined acceptance level, before the affected test sequence was terminated and the failure was reported.

6.3.4.5 Data generation

Although pseudo random generation of operands and other data should have been straight forward, some instruction set particularities complicated the task. To obtain deterministic results for some floating point operations, the operands, in both the memory and internal registers, had to be normalised. If this was not performed, the outcome from the computation would have been unspecified and different behaviours could be expected for the two models. Some of the data structures required alignment to even addresses when stored in memory and registers for correct operation, which also had to be handled explicitly in the testbench.

Purely random verification of a floating point unit is unlikely to find all errors. For example, there is only a small chance that the mantissas of two random operands will overlap and result in an actual calculation (Ref. 5). To overcome this limitation and to excite corner cases, we generated weighted pseudo random

operands in the testbench, such as small or large negative and positive numbers, and floating point zero.

6.3.4.6 Observability and reporting

To improve the possibility for detecting errors, the visibility of the internal state of the microprocessor was increased by periodic readout of all internal registers. The instruction set architecture conveniently allowed all internal register to be written to memory by issuing a single instruction. The purpose was to allow the testbench to catch errors located in flags etc., that might not propagate to the memory before being overwritten. Note that no changes in the black box models were required.

An important trade-off to be made is how many random instructions should be executed before the internal registers should be read out. By allowing a large number of instructions between register dumps, errors might be masked by subsequent instruction execution. By only executing a few instructions between register dumps, errors due to interactions between instructions might be masked and the efficiency decreases due to overhead. The testbench implementation allowed an arbitrary number of instructions to be executed between internal register dumps.

To assist debugging, the testbench generated a disassembled listing of the executed instruction sequence. Not all code was reported, only the last relevant part before a major deviation occurred between the test object and the reference. The disassembly was also performed in VHDL, using the instruction set data base. The listing further indicated what instructions or operands had been dropped or added to make up for the differences in model behaviour, what data deviations had occurred

and where exception handling had taken place in the form of a call to a interrupt routine.

A simple profile of the executed instructions was also created during the simulation, indicating how many times each instruction had been executed. More advance profiling tools have been reported (Ref.2, Ref.4),

6.3.4.7 Results

The verification resulted not only in the discovery of a large number of deviations between the board level model and the reference, but also revealed a few errors in existing hardware. Needless to say, the functional errors were corrected in the next revision of the microprocessor.

There were also several non-errors that were reported, requiring some effort to analyse and occasionally to adapt the testbench to cope with legal deviations between the models. After a couple of iterations, the testbench was capable of handling all acceptable deviations and long simulation runs could then have taken place. This was however not possible since new errors were regularly discovered which prevented any long instructions sequences, except for when all erroneous instructions were disabled.

The method will generally report many errors for each incorrect instruction implementation, which will require extensive analysis before resolving the cause behind each report. The approach is consequently suited for demonstrating that there are no or few errors left in a design at the end of the development.

Considering that the board level model had already been thoroughly verified, the selection of the approach was right at that time. If knowing that so many errors would be discovered,

perhaps a simpler approach could have been taken to eliminate the bulk of the problems before employing the presented approach to find the remaining errors at a lower expense than what had been the case.

6.3.5 Related Verification Cases

The program-less approach is not limited to VHDL, but can be applied to any hardware description language as long as it is possible to generate the stimuli on the fly, while taking into account deviating responses from the test object and the reference. The following two examples will illustrate how the approach has been used, or how it sometimes has not been possible to use due to surrounding constraints.

6.3.5.1 Microcontroller

To illustrate some limitations with off-line generation of pseudo random programs, we have included an example on how the discovery of an error caused lengthy analysis work due to development tool limitations. A microcontroller was developed based on a synthesisable VHDL core from an intellectual property provider.

The test object was first subjected to all verification suits that were provided with the core, followed by pseudo random testing. The reference was represented as a commercial hardware emulator that was part of the software development environment. The stimuli was generated with a custom made C program and was output as an executable binary. In total, three errors were discovered by this additional verification.

One error in the core was located in a jump instruction, resulting in an incorrect program counter update. Since the stimuli

was executed as a program, the jump introduced an infinite loop that never terminated. The simulation run for several hours before it was decided to abort it. Since the program did not terminate, there was no indication of what had gone wrong. To trace the problem, the stimuli was disassembled and run with a software debugger coupled with the VHDL simulator until the incorrect jump was identified. What further complicated the analysis process was that the available assembler was not able to generate the jump instruction under the same conditions as during the pseudo random test. A diagnostics program had to be assembled by hand before the cause of the error could be confirmed.

This example illustrates that pseudo random generators which rely on software tools are subject to a number of limitations. The problems with locating the error would have not occurred if the instructions would have been generated on the fly as in the program-less approach, since an infinite loop would not have been created. The execution would have terminated as soon as the number of deviating addresses issued during instruction fetching would have exceeded an arbitrary tolerance level.

6.3.5.2 Digital signal processor

A simplified version of the approach was applied during the validation of a digital signal processor. The device in question had been transferred from a commercial technology to a radiation-tolerant process. The random stimuli were applied to the first prototype devices and to the already existing commercial counterpart. The objective was to prove that the two devices were identical, rather than that the new device fulfilled all user requirements. In this setup, the applicability of the method was limited since program-less verification could not be applied without

redesigning the hardware. This was however compensated by the speed of the execution, which allowed pseudo random instruction suites with lengths widely superseding what is feasible to simulate.

The only discrepancy discovered between the reference and the test object was related to one single arithmetic operation, where one bit in a status register was flagged one clock period too late. It turned out that the reference was of an older revision of the processor than what had been used for developing the radiation-tolerant device. The error was thus located in the reference and not in the test object. In fact, the only documented difference between the two revisions was the anomaly in said arithmetic operation. This single anomaly caused the test program to report many errors, which indicates that the effectiveness of the verification approach was high.

6.3.6 Conclusions

A program-less approach to microprocessor verification using pseudo random stimuli generation eliminates many of the limitations associated with verification through program execution and with related software development tools.

As illustrated in the examples, the method will generally report many errors for each incorrect instruction implementation, which will require extensive analysis before resolving the cause behind each report. The method is therefore most suited for verifying that there are no errors in a design, rather than as a means for identifying errors during the device development.

For complex designs it is necessary to obtain a correct reference model that can be used throughout the development process. By describing the reference at a high abstraction level,

the functionality is not obscured by implementation details which reduces the possibility for introduction of errors. The difference in abstraction levels will however most likely result in behavioural differences.

Although the method tolerates behavioural differences between the test object and the reference, analysis of errors discovered by pseudo random stimuli can be very time consuming. A design should therefore be properly verified throughout the development phase, since errors detected late in the verification process are costly to handle (Ref.8). For a few of the discussed examples it was not possible to perform program-less stimuli generation that adapted to the response from the execution. The main reason for this was that the reference or test object was represented as existing hardware.

The need for a fast reference model when developing devices in excess of half a million gates has been claimed in several reports (Ref.2, Ref.5). In most cases the reference model is developed in the C language to offer high simulation performance as compared to VHDL. Our and others' experiences are that the simulation time requires a lesser part of the verification phase than what is needed for developing and debugging of testbenches (Ref. 1). The conclusion is that the speed of the reference model is not that important when the test object is modelled at a low abstraction level and will thus in any case be responsible for most of the simulation time.

References

- 1 Functional Verification of Large ASICs, A. Evans et al., Proceedings of the 35th DAC, 1998

- 2 I'm Done Simulating; Now What? Verification Coverage Analysis and Correctness Checking of the DECchip 21 164 Alpha Microprocessor, M. Kantrowita et al. , Proceedings of the 33rd DAC, 1996
- 3 Hierarchical Random Simulation Approach for the Verification of S/390 CMOS Multiprocessors, J. Walter et al. , Proceedings of the 34th DAC, 1997
- 4 Code Generation and Analysis for the Functional Verification of Microprocessors, A. Hosseini et al. , Proceedings of the 33rd DAC, 1996
- 5 Verifying the PA-8000's FPU, D. Smentek et al. , Integrated System Design Magazine, March 1997
- 6 Functional Verification Methodology of Chameleon Processor, F. Casaubieilh et al. , Proceedings of the 33rd DAC, 1996
- 7 Using VHDL for Board Level Simulation, S. Habinc, P. Sinander, IEEE Design and Test of Computers, Autumn 1996
- 8 A C-Based RTL Design Verification Methodology or Complex Microprocessors, J. Yim et al. , Proceedings of the 34th DAC, 1997

6.4 What Is Nanotechnology

6.4.1 Introduction

Despite unprecedented government funding and public interest in nanotechnology, few can accurately define the scope, range or potential applications of this technology. One of the most pressing issues facing nanoscientists and technologists today is that of communicating with the non-scientific community. As a result of decades of speculation, a number of myths have grown up around the field, making it difficult for the general public, or indeed the business and financial communities, to understand what is a fundamental shift in the way we look at our interactions with the

natural would. This article attempts to address some of these misconceptions, and explain why scientists, businesses and governments are spending large amounts of time and money on nanoscale research and development.

6.4.2 What Is Nanotechnology

Take a random selection of scientists, engineers, investors and the general public and ask them what nanotechnology is and you will receive a range of replies as broad as nanotechnology itself. For many scientists, it is nothing startlingly new; after all we have been working at the nanoscale for decades, through electron microscopy, scanning probe microscopies or simply growing and analysing thin films. For most other groups, however, nanotechnology means something far more ambitious, miniature submarines in the bloodstream little cogs and gears made out of atoms, space elevators made of nanotubes, and the colonization of space. It is no wonder people often muddle up nanotechnology with science fiction.

6.4.3 What Is the Nanoscale

Although a metre is defined by the International Standards Organization as ‘the length of the path travelled by light in vacuum during a time interval of $1/299\,792\,458$ of a second’ and a nanometre is by definition 10^{-9} of a metre, this does not help scientists to communicate the nanoscale to non-scientists. It is in human nature to relate sizes by reference to everyday objects. and the commonest definition of nanotechnology is in relation to the width of a human hair.

Unfortunately, human hairs are highly variable, ranging from

tens to hundreds of microns in diameter (10^{-6} of a metre), depending on the colour, type and the part of the body from which they are taken, so what is needed is a standard to which we can relate the nanoscale. Rather than asking anyone to imagine a millionth or a billionth of something, which few sane people can accomplish with ease, relating nanotechnology to atoms often makes the nanometre easier to imagine. While few non-scientists have a clear idea of how large an atom is, defining a nanometre as the size of 10 hydrogen, or 5 silicon atoms in a line is within the power of the human mind to grasp. The exact size of the atoms is less important than communicating the fact that nanotechnology is dealing with the smallest parts of matter that we can manipulate.

6.4.4 Science Fiction

While there is a commonly held belief that nanotechnology is a futuristic science with applications 25 years in the future and beyond, nanotechnology is anything but science fiction. In the last 15 years over a dozen Nobel prizes have been awarded in nanotechnology, from the development of the scanning probe microscope (SPM), to the discovery of fullerenes. According to CMP Cientifica, over 600 companies are currently active in nanotechnology, from small venture capital backed start-ups to some of the world's largest corporations such as IBM and Samsung. Governments and corporations worldwide have ploughed over \$ 4 billion into nanotechnology in the last year alone. Almost every university in the world has a nanotechnology department, or will have at least applied for the funding for one.

Even more significantly, there are companies applying nanotechnology to a variety of products we can already buy, such

as automobile parts, clothing and ski wax. Nanotechnology is already all around us if you know where to look.

The confusion arises in part because many people in the business world do not know where to look. Over the last decade, technology has become synonymous with computers, software and communications, whether the internet or mobile telephones. Many of the initial applications of nanotechnology are materials related, such as additives for plastics, nanocarbon particles for improved steels, coatings and improved catalysts for the petrochemical industry. All of these are technology based industries, maybe not new ones, but industries with multibillion dollar markets.

6.4.5 The Nanotechnology Industry

It is increasingly common to hear people referring to 'the nanotechnology industry', just like the software or mobile phone industries, but will such a thing ever exist? Many of the companies working with nanotechnology are simply applying our knowledge of the nanoscale to existing industries, whether it is improved drug delivery mechanisms for the pharmaceutical industry, or producing nanoclay particles for the plastics industry. In fact nanotechnology is an enabling technology rather than an industry in its own right. No one would ever describe Microsoft or Oracle as being part of the electricity industry, even though without electricity the software industry could not exist. Rather, nanotechnology is a fundamental understanding of how nature works at the atomic scale. New industries will be generated as a result of this understanding, just as the understanding of how electrons can be moved in a conductor by applying a potential difference led to electric lighting, the telephone, computing, the internet and many

other industries, all of which would not have been possible without it.

While it is possible to buy a packet of nanotechnology, a gram of nanotubes for example, it would have zero intrinsic value. The real value of the nanotubes would be in their application, whether within existing industry, or to enable the creation of a whole new one.

6.4.6 Fantastic Voyage

Shrinking machines down to the size where they can be inserted into the human body in order to detect and repair diseased cells is a popular idea of the benefits of nanotechnology, and one that even comes close to reality. Many companies are already in clinical trials for drug delivery mechanisms based on nanotechnology, but unfortunately none of them involve miniature submarines. It turns out that there are a whole range of more efficient ways that nanotechnology can enable better drug delivery without resorting to the use of nanomachines.

Just the concept of navigating ones way around the body at will does not bear serious scrutiny. Imagine attempting to go against the flow in an artery—it would be like swimming upstream in a fast flowing river, while boulders the size of houses, red and white blood cells, rained down on you. Current medical applications of nanotechnology are far more likely to involve improved delivery methods, such as pulmonary or epidermal methods to avoid having to pass through the stomach, encapsulation for both delivery and delayed release, and eventually the integration of detection with delivery, in order for drugs to be delivered exactly where they are needed, thus minimizing side

effects on healthy tissue and cells. As far as navigation goes, delivery will be by exactly the same method that the human body uses, going with the flow and 'dropping anchor' when the drug encounters its target.

6.4.7 Shrinking Stuff

Another common misconception is that nanotechnology is primarily concerned with making things smaller. This has been exacerbated by images of tiny bulls, and miniature guitars that can be strummed with the tip of an AFM, that while newsworthy, merely demonstrate our new found control of matter at the sub-micron scale. While almost the whole focus of micro-technologies has been on taking macro-scale devices such as transistors and mechanical systems and making them smaller, nanotechnology is more concerned with our ability to create from the bottom up. In electronics, there is a growing realization that with the end of the CMOS roadmap in sight at around 10 nm, combined with the uncertainly principal's limit of Von Neuman electronics at 2 nm, that merely making things smaller will not help us. Replacing CMOS transistors on a one for one basis with some type of nano device would have the effect of drastically increasing fabrication costs, while offering only a marginal improvement over current technologies.

However, nanotechnology offers us a way out of this technological and financial cul-de-sac by building devices from the bottom up. Techniques such as self assembly, perhaps assisted by templates created by nano imprint lithography, a notable European success, combined with our understanding of the workings of polymers and molecules such as Rotaxane at the nanoscale open

up a whole new host of possibilities. Whether it is avoiding Moore's second law by switching to plastic electronics, or using molecular electronics, our understanding of the behaviour of materials on the scale of small molecules allows a variety of alternative approaches, to produce smarter, cheaper devices. The new understandings will also allow us to design new architectures, with the end result that functionality will become a more valid measure of performance than transistor density or operations per second.

6.4.8 Nanotechnology Is New

It often comes as a surprise to learn that the Romans and Chinese were using nanoparticles thousands of years ago. Similarly, every time you light a match, fullerenes are produced. Degussa have been producing carbon black, the substance that makes car tyres black and improves the wear resistance of the rubber, since the 1920s. Of course they were not aware that they were using nanotechnology, and as they had no control over particle size, or even any knowledge of the nanoscale they were not using nanotechnology as currently defined.

What is new about nanotechnology is our ability to not only see, and manipulate matter on the nanoscale, but our understanding of atomic scale interactions.

6.4.9 Building Atom by Atom

One of the defining moments in nanotechnology came in 1989 when Don Eigler used a SPM to spell out the letters IBM in xenon atoms. For the first time we could put atoms exactly where we wanted them, even if keeping them there at much above absolute

zero proved to be a problem. While useful in aiding our understanding of the nanoworld, arranging atoms together one by one is unlikely to be of much use in industrial processes. Given that a Pentium 4 processor contains 42 million transistors, even simplifying the transistors to a cube of 100 atoms on each side would require 42×10^{12} operations, and that is before we start to consider the other material and devices needed in a functioning processor.

Of course we already have the ability to build things atom by atom, and on a very large scale; it is called physical chemistry, and has been in industrial use for over a century producing everything from nitrates to salt. To do this, we do not need any kind of tabletop assembler as in Star Trek, usually a few barrels of readily available precursor chemicals and maybe a catalyst are all that is required.

Compare this with the difficulty of producing anything organic atom by atom, a sausage for example. Everyone is familiar with the macroscale ingredients of a sausage, some meat, maybe some fat, cartilage or other kinds of tissue, even some bone, all encased in animal gut. Never mind, argue the proponents of assemblers, things are simpler at smaller scales.

Zooming down to the microscale we still have far more complexity than we would like to attempt to replicate, with cells, cytoplasm, mitochondria, chromosomes, ribosomes and many other highly complex items of natural engineering. Moving closer to the nanoscale, we still have to deal with nucleic acids, nucleotides, peptides and proteins, none of which we fully understand, or expect to even have the computing power to understand in the near future.

In terms of return on our investment, a farmyard containing a few pigs is a far more effective sausage machine than we could ever design, and has several other by-products such as hams and a highly effective waste disposal system. This serves to illustrate just how far we are away from being able to replicate nature.

6.4.10 Attack of the Killer Nanobots

In terms of capturing the public imagination, unleashing hordes of self-replicating devices that escape from the lab and attack anything in their path is always going to be popular. Unfortunately nature has already beaten us to it, by several hundred million years. Naturally occurring nanomachines, that can not only replicate and mutate as they do so in order to avoid our best attempts at eradication, but can also escape their hosts and travel with alarming ease through the atmosphere. No wonder that viruses are the most successful living organisms on the planet, with most of their 'machinery' being well into the nano realm. However, there are finite limits to the spread of such 'nanobots', usually determined by their ability, or lack thereof, of converting a sufficiently wide range of material needed for future expansion. Indeed, the immune systems of many species, while unable to completely neutralize viruses without side effects such as runny noses, are so effective in dealing with this type of threat as a result of the wide range of different technologies available to a large complex organism when confronted with a single purpose nano-sized one. For any threat from the nano world to become a danger, it would have to include far more intelligence and flexibility than we could possibly design into it.

Our understanding of genomics and proteomics is primitive

compared with that of nature, and is likely to remain that way for the foreseeable future. For anyone determined to worry about nanoscale threats to humanity should consider mutations in viruses such as HIV that would allow transmission via mosquitoes, or deadlier versions of the influenza virus, which deserve far more concern than anything nanotechnology may produce.

6.4.11 Conclusions

Nanotechnology, like any other branch of science, is primarily concerned with understanding how nature works. We have discussed how our efforts to produce devices and manipulate matter are still at a very primitive stage compared to nature. Nature has the ability to design highly energy efficient systems that operate precisely and without waste, fix only that which needs fixing, do only that which needs doing, and no more. We do not, although one day our understanding of nanoscale phenomena may allow us to replicate at least part of what nature accomplishes with ease.

While many branches of what now falls under the umbrella term nanotechnology are not new, it is the combination of existing technologies with our new found ability to observe and manipulate at the atomic scale that makes nanotechnology so compelling from scientific, business and political viewpoints.

For the scientist, advancing the sum total of human knowledge has long been the driving force behind discovery, from the gentleman scientists of the 17th and 18th centuries to our current academic infrastructure. Nanotechnology is at a very early stage in our attempts to understand the world around us, and will provide inspiration and drive for many generations of scientists.

For business, nanotechnology is no different from any other technology: it will be judged on its ability to make money. This may be in the lowering of production costs by, for example, the use of more efficient or more selective catalysts in the chemicals industry, by developing new products such as novel drug delivery mechanisms or stain resistant clothing, or the creation of entirely new markets, as the understanding of polymers did for the multi-billion euro plastics industry.

Politically, it can be argued that fear is the primary motivation. The US has opened up a commanding lead in terms of economic growth, despite recent setbacks, as a result of the growth and adoption of information technology. Of equal significance is the lead in military technology as demonstrated by the use of unmanned drones for both surveillance and assault in recent conflicts. Nanotechnology promises far more significant economic, military and cultural changes than those created by the internet, and with technology advancing so fast, and development and adoption cycles becoming shorter, playing catch-up will not be an option for governments who are not already taking action.

Maybe the greatest short term benefit of nanotechnology is in bringing together the disparate sciences, physical and biological, who due to the nature of education often have had no contact since high school. Rather than nanosubmarines or killer nanobots, the greatest legacy of nanotechnology may well prove to be the unification of scientific disciplines and the resultant ability of scientists, when faced with a problem, to call on the resources of the whole of science, not just of one discipline.

参 考 文 献

- 1 Yangyuan Wang , Ruqi Han, Xiaoyuan Liu. The Challenges for Physical Limitation in Si Microelectronics. 1998 5th International Conference on Solis-State and Integrated Circuit Technology, 1998, Beijing, CHINA, 25 ~ 30

Images have been losslessly embedded. Information about the original file can be found in PDF attachments. Some stats (more in the PDF attachments):

```
{
  "filename": "MTExMjAyMjkuemlw",
  "filename_decoded": "11120229.zip",
  "filesize": 25644044,
  "md5": "8f38a9154acec2c4a54a17c6924ab7fc",
  "header_md5": "977b88ee0e1d81c13298acecfd31e644",
  "sha1": "a64c14f0c0eeebb966f68d46d7d1464db32cb653",
  "sha256": "7395f05b8e2f162080cb403c220f8730fe2b1ec9c20c6df1ec91f603f0f40bcd",
  "crc32": 53071422,
  "zip_password": "52gv",
  "uncompressed_size": 26546571,
  "pdg_dir_name": "\u00ed\u2562\u2561\u03c4\u256b\u2559\u2510\u255e\u2564\u00ba\u2559\u03b4\u255d\u255d\u2569\u2321\u256b\u00bf\u2565\u2561\u2559\u00f3\u2559\u2229\u00ed\u00f1\u256c\u00f3\u2561\u03c4\u256b\u2559\u255d\u255d\u2569\u2321\u2556\u2553\u2593\u00df\u00ed\u2556_11120229",
  "pdg_main_pages_found": 401,
  "pdg_main_pages_max": 401,
  "total_pages": 409,
  "total_pixels": 1560487552,
  "pdf_generation_missing_pages": false
}
```