

PHYSICS OF SEMICONDUCTOR DEVICES



T H I R D E D I T I O N

S. M. SZE
KWOK K. NG

Physics of Semiconductor Devices

Third Edition

S. M. Sze

Department of Electronics Engineering
National Chiao Tung University
Hsinchu, Taiwan

and

Kwok K. Ng

Central Laboratory
MVC (a subsidiary of ProMOS Technologies, Taiwan)
San Jose, California



A JOHN WILEY & SONS, INC., PUBLICATION

Description of cover photograph:

A scanning electron micrograph of an array of the floating-gate nonvolatile semiconductor memory (NVSM) magnified 100,000 times. NVSM was invented at Bell Telephone Laboratories in 1967. There are more NVSM cells produced annually in the world than any other semiconductor device and, for that matter, any other human-made item. For a discussion of this device, see Chapter 6. Photo courtesy of Macronix International Company, Hsinchu, Taiwan, ROC.

Copyright © 2007 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permission>.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic format. For information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data is available.

ISBN-13: 978-0-471-14323-9
ISBN-10: 0-471-14323-5

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

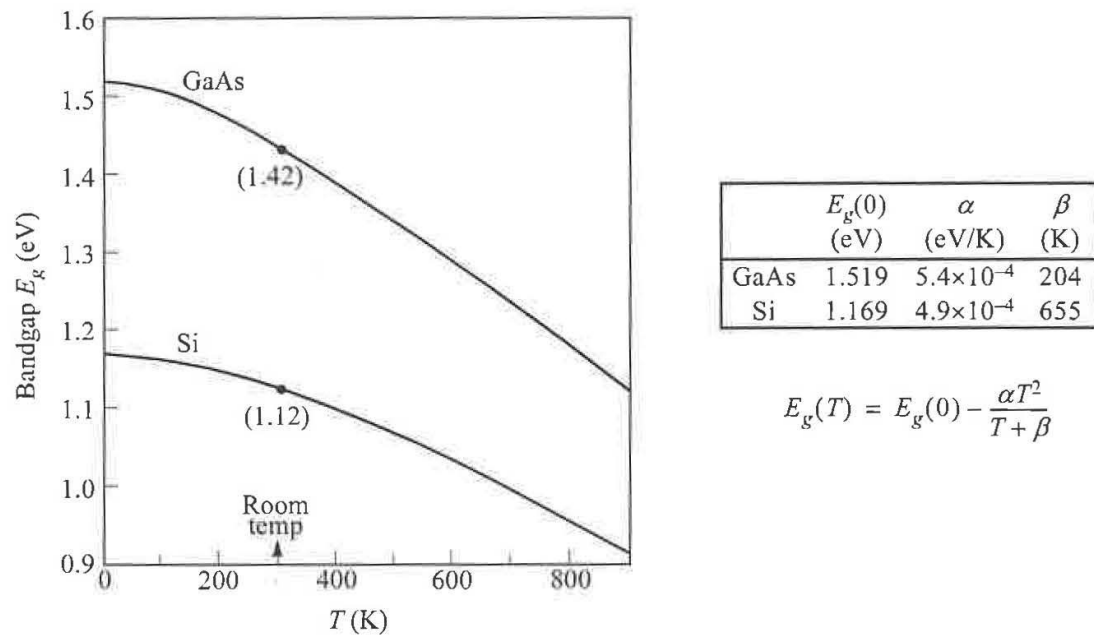


Fig. 6 Energy bandgaps of Si and GaAs as a function of temperature. (After Refs. 22–23.)

1.4 CARRIER CONCENTRATION AT THERMAL EQUILIBRIUM

One of the most-important properties of a semiconductor is that it can be doped with different types and concentrations of impurities to vary its resistivity. Also, when these impurities are ionized and the carriers are depleted, they leave behind a charge density that results in an electric field and sometimes a potential barrier inside the semiconductor. Such properties are absent in a metal or an insulator.

Figure 7 shows three basic bond representations of a semiconductor. Figure 7a shows intrinsic silicon, which is very pure and contains a negligibly small amount of impurities. Each silicon atom shares its four valence electrons with the four neigh-

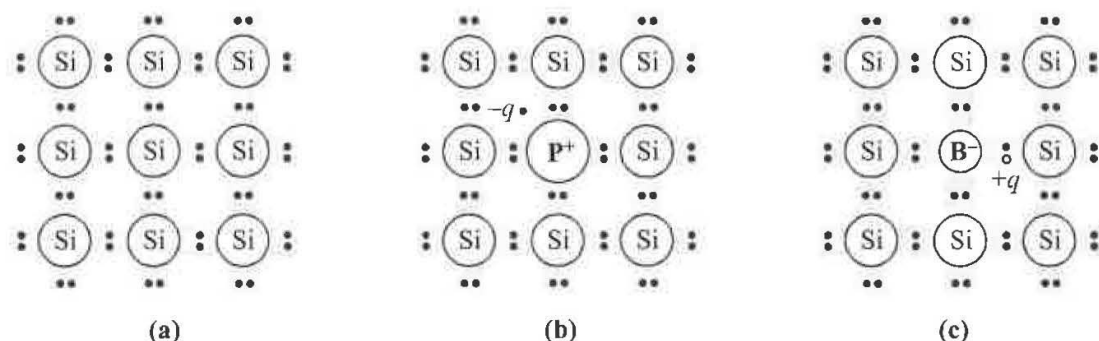


Fig. 7 Three basic bond pictures of a semiconductor. (a) Intrinsic Si with no impurity. (b) n -type Si with donor (phosphorus). (c) p -type Si with acceptor (boron).

boring atoms, forming four covalent bonds (also see Fig. 1). Figure 7b shows an *n*-type silicon, where a substitutional phosphorous atom with five valence electrons has replaced a silicon atom, and a *negative*-charged electron is *donated* to the lattice in the conduction band. The phosphorous atom is called a *donor*. Figure 7c similarly shows that when a boron atom with three valence electrons substitutes for a silicon atom, a *positive*-charged *hole* is created in the valence band, and an additional electron will be *accepted* to form four covalent bonds around the boron. This is *p*-type, and the boron is an *acceptor*.

These names of *n*- and *p*-type had been coined when it was observed that if a metal whisker was pressed against a *p*-type material, forming a Schottky barrier diode (see Chapter 3), a *positive* bias was required on the semiconductor to produce a noticeable current.^{25,26} Also, when exposed to light, a *positive* potential was generated with respect to the metal whisker. Conversely, a *negative* bias was required on an *n*-type material to produce a large current.

1.4.1 Carrier Concentration and Fermi Level

We first consider the intrinsic case without impurities added to the semiconductor. The number of electrons (occupied conduction-band levels) is given by the total number of states $N(E)$ multiplied by the occupancy $F(E)$, integrated over the conduction band,

$$n = \int_{E_C}^{\infty} N(E)F(E)dE. \quad (13)$$

The density of states $N(E)$ can be approximated by the density near the bottom of the conduction band for low-enough carrier densities and temperatures:⁵

$$N(E) = M_C \frac{\sqrt{2} m_{de}^{3/2} (E - E_C)^{1/2}}{\pi^2 \hbar^3}. \quad (14)$$

M_C is the number of equivalent minima in the conduction band and m_{de} is the density-of-state effective mass for electrons:⁵

$$m_{de} = (m_1^* m_2^* m_3^*)^{1/3} \quad (15)$$

where m_1^* , m_2^* , m_3^* are the effective masses along the principal axes of the ellipsoidal energy surface. For example, in silicon $m_{de} = (m_l^* m_t^{*2})^{1/3}$. The occupancy is a strong function of temperature and energy, and is represented by the Fermi-Dirac distribution function

$$F(E) = \frac{1}{1 + \exp[(E - E_F)/kT]} \quad (16)$$

where E_F is the Fermi energy level which can be determined from the charge neutrality condition (see Section 1.4.3).

The integral of Eq. 13 can be evaluated to be

$$n = N_C \frac{2}{\sqrt{\pi}} F_{1/2} \left(\frac{E_F - E_C}{kT} \right) \quad (17)$$

where N_C is the effective density of states in the conduction band and is given by

$$N_C \equiv 2 \left(\frac{2 \pi m_{de} kT}{h^2} \right)^{3/2} M_C. \quad (18)$$

The Fermi-Dirac integral, changing variables with $\eta \equiv (E - E_C)/kT$ and $\eta_F \equiv (E_F - E_C)/kT$, is given by

$$\begin{aligned} F_{1/2} \left(\frac{E_F - E_C}{kT} \right) &\equiv F_{1/2}(\eta_F) = \int_{E_C}^{\infty} \frac{[(E - E_C)/kT]^{1/2} dE}{1 + \exp[(E - E_F)/kT]} \\ &= \int_0^{\infty} \frac{\eta^{1/2}}{1 + \exp(\eta - \eta_F)} d\eta \end{aligned} \quad (19)$$

whose values are plotted in Fig. 8. Note that for $\eta_F < -1$, the integral can be approximated by an exponential function. At $\eta_F = 0$ when the Fermi level coincides with the band edge, the integral has a value of ≈ 0.6 such that $n \approx 0.7N_C$.

Nondegenerate Semiconductors. By definition, in nondegenerate semiconductors, the doping concentrations are smaller than N_C and the Fermi levels are more than several kT below E_C (negative η_F), the Fermi-Dirac integral approaches

$$F_{1/2} \left(\frac{E_F - E_C}{kT} \right) = \frac{\sqrt{\pi}}{2} \exp \left(- \frac{E_C - E_F}{kT} \right) \quad (20)$$

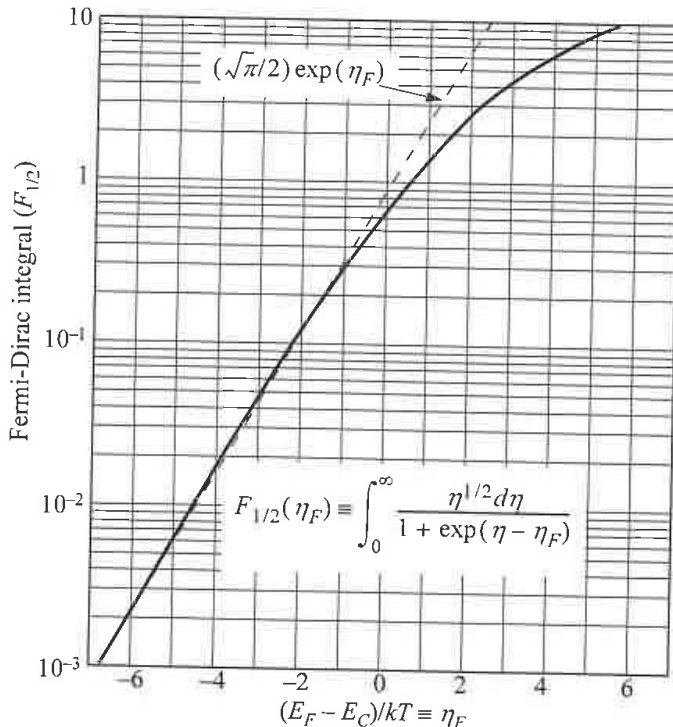


Fig. 8 Fermi-Dirac integral $F_{1/2}$ as a function of Fermi energy. (After Ref. 27.) Dashed line is approximation of Boltzmann statistics.

and Boltzmann statistics apply. Equation 17 becomes

$$n = N_C \exp\left(-\frac{E_C - E_F}{kT}\right) \quad \text{or} \quad E_C - E_F = kT \ln\left(\frac{N_C}{n}\right). \quad (21)$$

Similarly, for p -type semiconductors we can obtain the hole density and its Fermi level near the top of the valence band;

$$p = N_V \frac{2}{\sqrt{\pi}} F_{1/2}\left(\frac{E_V - E_F}{kT}\right) \quad (22)$$

which can be simplified to

$$p = N_V \exp\left(-\frac{E_F - E_V}{kT}\right) \quad \text{or} \quad E_F - E_V = kT \ln\left(\frac{N_V}{p}\right), \quad (23)$$

where N_V is the effective density of states in the valence band and is given by

$$N_V \equiv 2 \left(\frac{2\pi m_{dh} kT}{h^2} \right)^{3/2}. \quad (24)$$

Here m_{dh} is the density-of-state effective mass of the valence band:⁵

$$m_{dh} = (m_{lh}^{*3/2} + m_{hh}^{*3/2})^{2/3} \quad (25)$$

where the subscripts refer to *light* and *heavy* hole masses previously referenced in Eq. 9.

Degenerate Semiconductors. As shown in Fig. 8, for degenerate levels where n - or p -concentrations are near or beyond the effective density of states (N_C or N_V), the value of Fermi-Dirac integral has to be used instead of the simplified Boltzmann statistics. For $\eta_F > -1$, the integral has weaker dependence on the carrier concentration. Note that also the Fermi levels are outside the energy gap. A useful estimate of the Fermi level as a function of carrier concentration is given by, for n -type semiconductor²⁸

$$E_F - E_C \approx kT \left[\ln\left(\frac{n}{N_C}\right) + 2^{-3/2} \left(\frac{n}{N_C}\right) \right], \quad (26a)$$

and for p -type

$$E_V - E_F \approx kT \left[\ln\left(\frac{p}{N_V}\right) + 2^{-3/2} \left(\frac{p}{N_V}\right) \right]. \quad (26b)$$

Intrinsic Concentration. For intrinsic semiconductors at finite temperatures, thermal agitation occurs which results in continuous excitation of electrons from the valence band to the conduction band, and leaving an equal number of holes in the valence band. This process is balanced by recombination of the electrons in the conduction band with holes in the valence band. At steady state, the net result is $n = p = n_i$, where n_i is the intrinsic carrier density.

The Fermi level for an intrinsic semiconductor (which by definition is nondegenerate) is obtained by equating Eqs. 21 and 23:

$$\begin{aligned}
 pn &= N_C N_V \exp\left(-\frac{E_g}{kT}\right) \\
 &= n_i^2,
 \end{aligned} \tag{29}$$

which is known as the mass-action law. But for degenerate semiconductors, $pn < n_i^2$. Also using Eq. 28 and E_i as the reference energy, we have the alternate equations for n -type materials;

$$n = n_i \exp\left(\frac{E_F - E_i}{kT}\right) \quad \text{or} \quad E_F - E_i = kT \ln\left(\frac{n}{n_i}\right), \tag{30a}$$

and for p -type materials;

$$p = n_i \exp\left(\frac{E_i - E_F}{kT}\right) \quad \text{or} \quad E_i - E_F = kT \ln\left(\frac{p}{n_i}\right). \tag{30b}$$

1.4.2 Donors and Acceptors

When a semiconductor is doped with donor or acceptor impurities, impurity energy levels are introduced that usually lie within the energy gap. A donor impurity has a donor level which is defined as being neutral if filled by an electron, and positive if empty. Conversely, an acceptor level is neutral if empty and negative if filled by an electron. These energy levels are important in calculating the fraction of dopants being ionized, or electrically active, as discussed in Section 1.4.3.

To get a feeling of the magnitude of the impurity ionization energy, we use the simplest calculation based on the hydrogen-atom model. The ionization energy for the hydrogen atom in vacuum is

$$E_H = \frac{m_0 q^4}{32 \pi^2 \epsilon_0^2 \hbar^2} = 13.6 \text{ eV}. \tag{31}$$

The ionization energy for a donor ($E_C - E_D$) in a lattice can be obtained by replacing m_0 by the conductivity effective mass of electrons⁵

$$m_{ce} = 3 \left(\frac{1}{m_1^*} + \frac{1}{m_2^*} + \frac{1}{m_3^*} \right)^{-1} \tag{32}$$

and by replacing ϵ_0 by the permittivity of the semiconductor ϵ_s in Eq. 31:

$$E_C - E_D = \left(\frac{\epsilon_0}{\epsilon_s} \right)^2 \left(\frac{m_{ce}}{m_0} \right) E_H. \tag{33}$$

The ionization energy for donors as calculated from Eq. 33 is 0.025 eV for Si and 0.007 eV for GaAs. The hydrogen-atom calculation for the ionization level for the acceptors is similar to that for the donors. The calculated acceptor ionization energy (measured from the valence-band edge, $E_a \equiv (E_A - E_V)$) is 0.05 eV for Si and GaAs.

Although this simple hydrogen-atom model given above certainly cannot account for the details of ionization energy, particularly the deep levels in semiconductors,³¹⁻³³ the calculated values do predict the correct order of magnitude of the true ionization energies for shallow impurities. These calculated values are shown to be

much smaller than the energy gap, and often are referred to as shallow impurities if they are close to the band edges. Also, since these small ionization energies are comparable to the thermal energy kT , ionization is usually complete at room temperature. Figure 10 shows the measured ionization energies for various impurities in Si and GaAs. Note that it is possible for a single atom to have many levels; for example, gold in silicon has both an acceptor level and a donor level in the forbidden energy gap.

1.4.3 Calculation of Fermi Level

The Fermi level for the intrinsic semiconductor (Eq. 27) lies very close to the middle of the bandgap. Figure 11a depicts this situation, showing schematically from left to right the simplified band diagram, the density of states $N(E)$, the Fermi-Dirac distribution function $F(E)$, and the carrier concentrations. The shaded areas in the conduction band and the valence band represent electrons and holes, and their numbers are the same; i.e., $n = p = n_i$ for the intrinsic case.

When impurities are introduced to the semiconductor crystals, depending on the impurity energy level and the lattice temperature, not all dopants are necessarily ionized. The ionized concentration for donors is given by³⁶

$$N_D^+ = \frac{N_D}{1 + g_D \exp[(E_F - E_D)/kT]} \quad (34)$$

where g_D is the ground-state degeneracy of the donor impurity level and equal to 2 because a donor level can accept one electron with either spin (or can have no electron). When acceptor impurities of concentration N_A are added to a semiconductor crystal, a similar expression can be written for the ionized acceptors

$$N_A^- = \frac{N_A}{1 + g_A \exp[(E_A - E_F)/kT]} \quad (35)$$

where the ground-state degeneracy factor g_A is 4 for acceptor levels. The value is 4 because in most semiconductors each acceptor impurity level can accept one hole of either spin and the impurity level is doubly degenerate as a result of the two degenerate valence bands at $k = 0$.

When impurity atoms are introduced, the total negative charges (electrons and ionized acceptors) must equal the total positive charges (holes and ionized donors), represented by the charge neutrality

$$n + N_A^- = p + N_D^+. \quad (36)$$

With impurities added, the mass-action law ($pn = n_i^2$) in Eq. 29 still applies (until degeneracy), and the pn product is always independent of the added impurities.

Consider the case shown in Fig. 11b, where donor impurities with the concentration N_D (cm^{-3}) are added to the crystal. The charge neutrality condition becomes

$$\begin{aligned} n &= N_D^+ + p \\ &\approx N_D^+ \end{aligned} \quad (37)$$

With substitution, we obtain

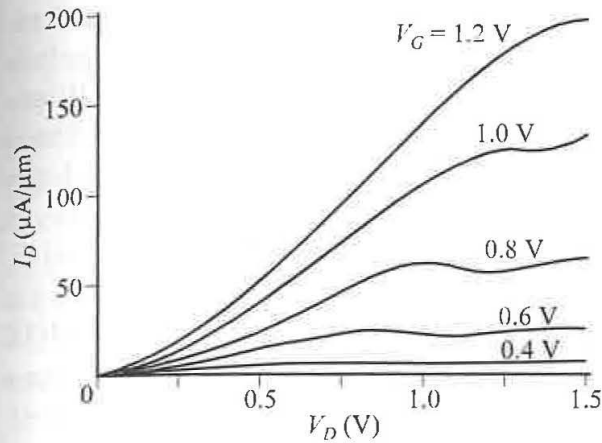


Fig. 35 I - V characteristics of n -channel MOSFET with Schottky source/drain. (After Ref. 63.)

The disadvantages of the Schottky source/drain are high series resistance due to the finite barrier height, and higher drain leakage current. Typical I - V curves show that current is starved at low-drain bias (Fig. 35). Also note that as shown in Fig. 34, the metal or silicide contact has to extend underneath the gate for continuity. This process is much more demanding than a junction source/drain which is done by self-aligned implantation and diffusion.

Raised Source/Drain. An advanced design is the raised source/drain where a heavily doped epitaxial layer is grown over the source/drain regions (Fig. 36a). The purpose is to minimize junction depth to control short-channel effects. Note that an extension underneath the spacer is still needed for continuity. An alternative is the recessed-channel MOSFET where the junction depth r_j is zero or negative (Fig. 36b).⁶⁴ The drawback of the recessed-channel structure, especially for submicron devices, is the difficulty in controlling the contour and the oxide thickness at the corners where the threshold voltage is determined. Also, oxide charging may be worsened because more hot-carrier injection will occur.

6.5.4 SOI and Thin-Film Transistor (TFT)

SOI. Unlike thin-film transistor, the top silicon layer of an SOI (silicon-on-insulator) wafer is high-quality single-crystalline material that is suitable for high-performance

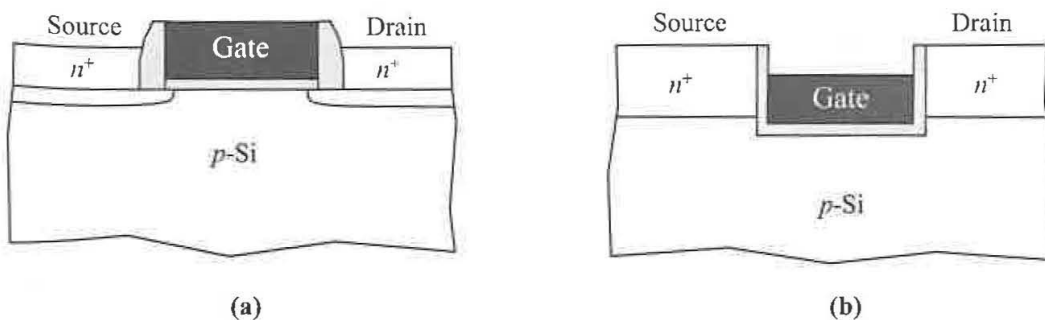


Fig. 36 Means to reduce source/drain junction depth and series resistance. (a) Raised source/drain. (b) Recessed channel.

and high-density integrated circuits.⁶⁵ Many forms of SOI structures have been demonstrated with different insulator materials and holding substrates. These include silicon-on-oxide, silicon-on-sapphire (SOS), silicon-on-zirconia (SOZ), and silicon-on-nothing (air gap). In SOS and SOZ technologies, a single-crystalline silicon film is epitaxially grown on a crystalline insulating substrate. In these cases, the insulators are the substrates themselves, Al_2O_3 in SOS and ZrO_2 in SOZ. The difficulties in these techniques are the material quality when the film gets thinner. The first option, using oxide as an insulator and another Si wafer as the holding substrate, is by far the most popular. There are many ways to fabricate this structure. Among them SIMOX (separation by implantation of oxygen) where high-dose oxygen is implanted onto a silicon wafer followed by high-temperature anneal to form the buried SiO_2 layer. Another technique involves bonding of two silicon wafers one of which has an oxidized layer followed by thinning or removal of the majority of the top wafer until a thin silicon layer is left. One technique uses lateral epitaxial growth of silicon over an oxide layer, starting from a seed opening to the substrate. Another uses laser recrystallization, transforming amorphous silicon deposited onto the oxide layer into single-crystalline material, or into poly-crystalline form with large grain size.

Figure 37a shows a schematic diagram of an n -channel MOSFET made on an SOI substrate, with its typical I - V characteristics shown in Fig. 37b. The kinks associated with floating body without a substrate tie are noticeable.

The advantages of the SOI substrate include improved MOSFET scaling due to its thin body. A thin body can alleviate most problem with punch-through such that the channel can be doped lightly. The subthreshold swing is known to be improved. The buried oxide layer serves as good isolation to reduce capacitance to the substrate, giving rise to higher speed. As shown in Fig. 37a, device isolation is much easier, simply by removing the surrounding thin film. This can significantly improve the circuit density. This type of isolation, as opposed to junction isolation in planar technology, also eliminates latch-up phenomenon in CMOS circuits. The disadvantages of SOI is higher wafer cost, potentially inferior material properties, the kink effect, and worse heat conduction because of the oxide layer.

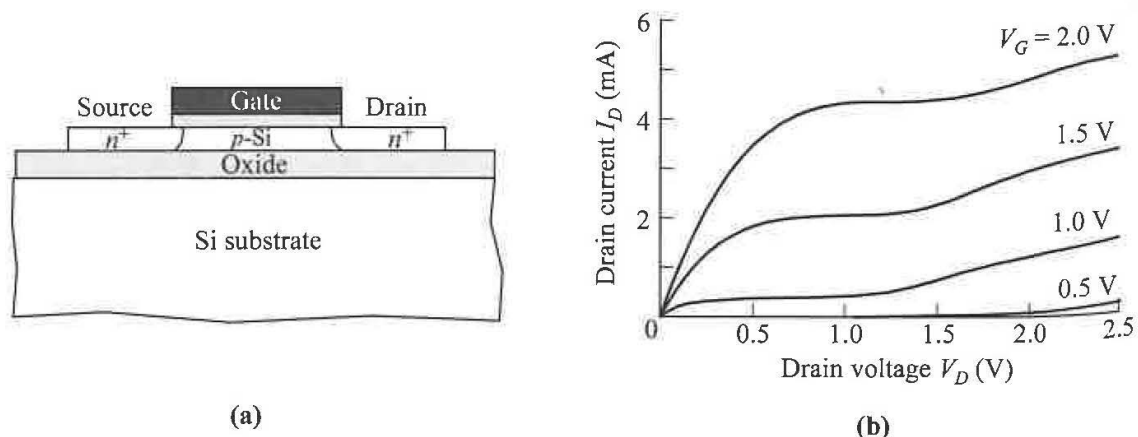


Fig. 37 (a) Typical structure of MOSFET on SOI wafer, and (b) its drain characteristics. (After Ref. 66.)

tors are cooled to lower temperatures (such as 77 K and 4.2 K). The lower temperatures reduce thermal effects which cause thermal ionization and deplete the energy levels, and increase the gain and detection efficiency. Near 0.5 μm , a CdS photoconductor has high sensitivity, whereas at 10 μm a HgCdTe photoconductor is preferred.⁶ In the wavelength range from 100 to 400 μm , a GaAs extrinsic photoconductor is a better choice because of its high detectivity.⁷ This photoconductor has high dynamic range and can give comparable performance for high-level (strong light intensity) detection. For low-level detection at microwave frequencies, however, a photodiode will provide considerably more speed and higher signal-to-noise ratio. Thus photoconductors have limited use in high-frequency optical demodulators, such as in optical mixing. They have been, however, extensively used for infrared detection especially beyond a few microns of wavelength.

13.3 PHOTODIODES

13.3.1 General Consideration

A photodiode has a depleted semiconductor region with a high electric field that serves to separate photogenerated electron-hole pairs. For high-speed operation, the depletion region must be kept thin to reduce the transit time. On the other hand, to increase the quantum efficiency (the number of electron-hole pairs generated per incident photon), the depletion layer must be sufficiently thick to allow a large fraction of the incident light to be absorbed. Thus, there is a trade-off between the speed of response and quantum efficiency.

For the visible and near-infrared wavelength range, photodiodes are usually reverse-biased with moderate biasing voltages, because this reduces the carrier transit time and lowers the diode capacitance. The reverse voltage is, however, not large enough to cause avalanche multiplication or breakdown. This biasing condition is in contrast to avalanche photodiodes, where an internal current gain is obtained as a result of the impact ionization under avalanche breakdown conditions. All photodiodes, with the exception of the avalanche photodiode which is not included in this section, thus have a maximum gain of one (Table 1). The photodiode family includes the *p-i-n* photodiode, *p-n* photodiode, heterojunction photodiode, and metal-semiconductor (Schottky barrier) photodiode.

We shall now briefly consider the general characteristics of a photodiode: its quantum efficiency, response speed, and device noise.

Quantum Efficiency. As mentioned previously, quantum efficiency is the number of electron-hole pairs generated per incident photon (Eq. 2). A related figure of merit is the responsivity, which is the ratio of the photocurrent to the optical power (Eq. 3). Therefore, for a given quantum efficiency, the responsivity increases linearly with wavelength. For an ideal photodiode ($\eta = 1$), $\mathcal{R} = \lambda/1.24$ (A/W) where λ is expressed in microns.

Since the optical absorption coefficient α is a strong function of the wavelength, for a given semiconductor the wavelength range in which appreciable photocurrent

can be generated is limited. Since most photodiodes use band-to-band photoexcitation (except for photoexcitation over the barrier in metal-semiconductor photodiodes), the long-wavelength cutoff λ_c is established by the energy gap of the semiconductor, Eq. 1, for example about 1.7 μm for Ge and 1.1 μm for Si. For wavelengths longer than λ_c , the values of α are too small to give appreciable absorption. The short-wavelength cutoff of the photoresponse comes about because the values of α are very large ($\geq 10^5 \text{ cm}^{-1}$), and the radiation is absorbed very near the surface where recombination is more likely. The photocarriers thus recombine before they are collected in the p - n junction. In the near-infrared region, silicon photodiodes with antireflection coating can reach 100% quantum efficiency near 0.8 to 0.9 μm . In the 1.0- to 1.6- μm region, Ge photodiodes, III-V ternary photodiodes (e.g., InGaAs), and III-V quaternary photodiodes (e.g., InGaAsP) have shown high quantum efficiencies. For longer wavelengths, photodiodes are cooled (e.g., 77 K) for high-efficiency operation.

Response Speed. The response speed is limited by a combination of three factors: (1) drift time in the depletion region, (2) diffusion of carriers, and (3) capacitance of the depletion region. Carriers generated outside the depletion region must diffuse to the junction resulting in considerable time delay. To minimize the diffusion effect, the junction should be formed very close to the surface. Most light will be absorbed when the depletion region is sufficiently wide (of the order of $1/\alpha$); with sufficient reverse bias the carriers will drift at their saturation velocities. The depletion layer must not be too wide, however, or transit-time effects will limit the frequency response. It also should not be too thin, or excessive capacitance C will result in a large $R_L C$ time constant, where R_L is the load resistance. The optimum compromise occurs when the depletion layer is chosen so that the transit time is of the order of one-half the modulation period. For example, for a modulation frequency of 10 GHz, the optimum depletion layer thickness in Si (with a saturation velocity of 10^7 cm/s) is about 5 μm .

Device Noise. To study the noise properties in a photodiode, we will consider the generalized photodetection process shown in Fig. 5a. An optical signal and background radiation are absorbed by the photodiode, whereby electron-hole pairs are generated. These electrons and holes are then separated by the electric field and drift toward the opposite sides of the junction. In the process, a photocurrent is induced in the external load resistor. Since noise is frequency dependent, to determine the currents generated by this photoelectric process, we will consider an intensity-modulated optical signal given by Eq. 12. The average photocurrent I_p due to the optical signal is given by Eq. 10. For the modulated optical signal, the rms signal power is $mP_{\text{opt}}/\sqrt{2}$, and the rms signal current is obtained from Eq. 13 with the gain set to unity,

$$i_p = \frac{q \eta m P_{\text{opt}}}{\sqrt{2} h \nu} \quad (17)$$

We designate the current resulting from the background radiation to be I_B , and the dark current due to thermal generation of electron-hole pairs in the depletion region

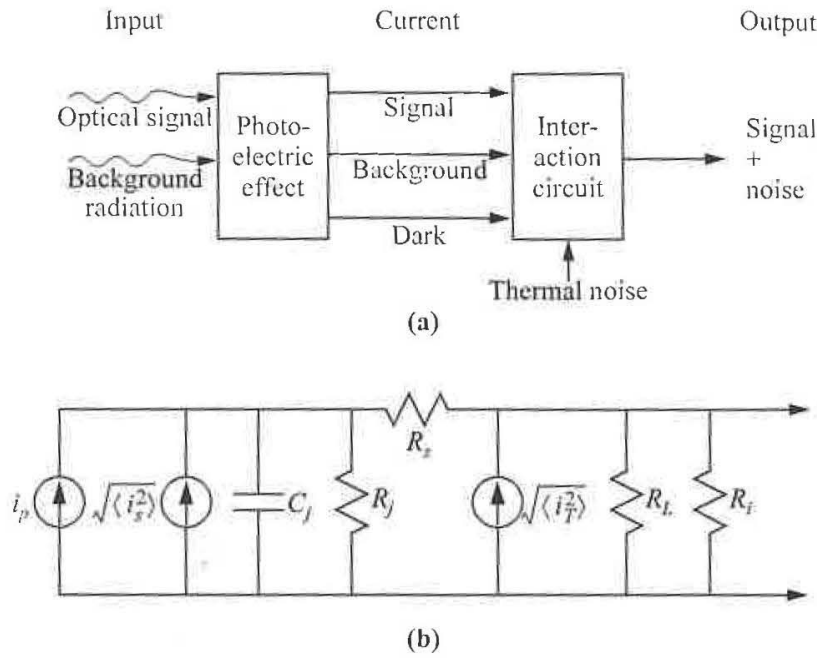


Fig. 5 Noise analysis of photodiode. (a) Photodetection process. (b) Equivalent circuit. (After Ref. 8.)

I_D . Because of the random generation of all these currents, they contribute to shot noise given by

$$\langle i_s^2 \rangle = 2q(I_P + I_B + I_D)B \quad (18)$$

where B is the bandwidth. The thermal noise is given by

$$\langle i_T^2 \rangle = \frac{4kTB}{R_{eq}} \quad (19)$$

where

$$\frac{1}{R_{eq}} = \frac{1}{R_j} + \frac{1}{R_L} + \frac{1}{R_i}. \quad (20)$$

The equivalent circuit of a photodiode is shown in Fig. 5b. The component C_j is the junction capacitance, R_j the junction resistance, and R_s the series resistance. The variable R_L is an external load resistor and R_i is the input resistance of the following amplifier.⁹ All the resistances contribute additional thermal noise to the system. The series resistance R_s is usually much smaller than the other resistances and can be neglected.

For a 100% modulated signal ($m = 1$) with average power P_{opt} , the signal-to-noise ratio can be written as

$$\left. \frac{S}{N} \right|_{\text{power}} = \frac{i_p^2}{\langle i_s^2 \rangle + \langle i_T^2 \rangle} = \frac{(1/2)(q\eta P_{opt}/h\nu)^2}{2q(I_P + I_B + I_D)B + 4kTB/R_{eq}}. \quad (21)$$

From this equation, the minimum optical power required to obtain a given signal-to-noise ratio is (setting $I_P = 0$)

$$P_{\text{opt}}|_{\min} = \frac{2h\nu}{\eta} \sqrt{\frac{(S/N)I_{eq}B}{q}} \quad (22)$$

where

$$I_{eq} = I_B + I_D + \frac{2kT}{qR_{eq}}. \quad (23)$$

The noise-equivalent power (NEP) is given by ($S/N = 1$, $B = 1$ Hz)

$$\begin{aligned} \text{NEP} &= \text{rms optical power } P_{\text{opt}}|_{\min} \\ &= \left(\frac{h\nu}{\eta}\right) \sqrt{\frac{2I_{eq}}{q}} \quad \text{W/cm}^2\text{-Hz}^{1/2}. \end{aligned} \quad (24)$$

To improve the sensitivity of a photodiode, both η and R_{eq} should be increased while I_B and I_D should be minimized. The NEP decreases with R_{eq} until it saturates to a constant value limited by dark-current or background-current shot noise.

13.3.2 *p-i-n* and *p-n* Photodiodes

The *p-i-n* photodiode is a special case of the *p-n* junction photodiodes, and is one of the most-common photodetectors, because the depletion-region thickness (the intrinsic layer) can be tailored to optimize the quantum efficiency and frequency response. Figure 6 shows schematic representation of a *p-i-n* diode and its energy-band diagram under reverse-bias conditions, together with optical absorption characteristics. We shall discuss the operation of *p-i-n* photodiode in some detail with the help of Fig. 6. This discussion applies also to *p-n* junction photodiodes. Light absorption in the semiconductor produces electron-hole pairs. Pairs produced in the depletion region or within a diffusion length of it will eventually be separated by the electric field, leading to current flow in the external circuit as carriers drift across the depletion layer.

Quantum Efficiency. Under steady-state conditions the total photocurrent density through the reverse-biased depletion layer is given by¹⁰

$$J_{\text{tot}} = J_{dr} + J_{\text{diff}} \quad (25)$$

where J_{dr} is the drift current due to carriers generated within the depletion region and J_{diff} is the diffusion current due to carriers generated outside the depletion layer in the bulk of the semiconductor and diffusing into the reverse-biased junction. We shall now derive the total current under the assumptions that the thermal generation current can be neglected and that the surface *p*-layer is much thinner than $1/\alpha$. Referring to Fig. 6c, the electron-hole generation rate is given by

$$G_e(x) = \Phi_0 \alpha \exp(-\alpha x) \quad (26)$$

where Φ_0 is the incident photon flux per unit area given by $P_{\text{opt}}(1 - R)/A h \nu$; R is the reflection coefficient, and A is the device area. The drift current J_{dr} is thus given by

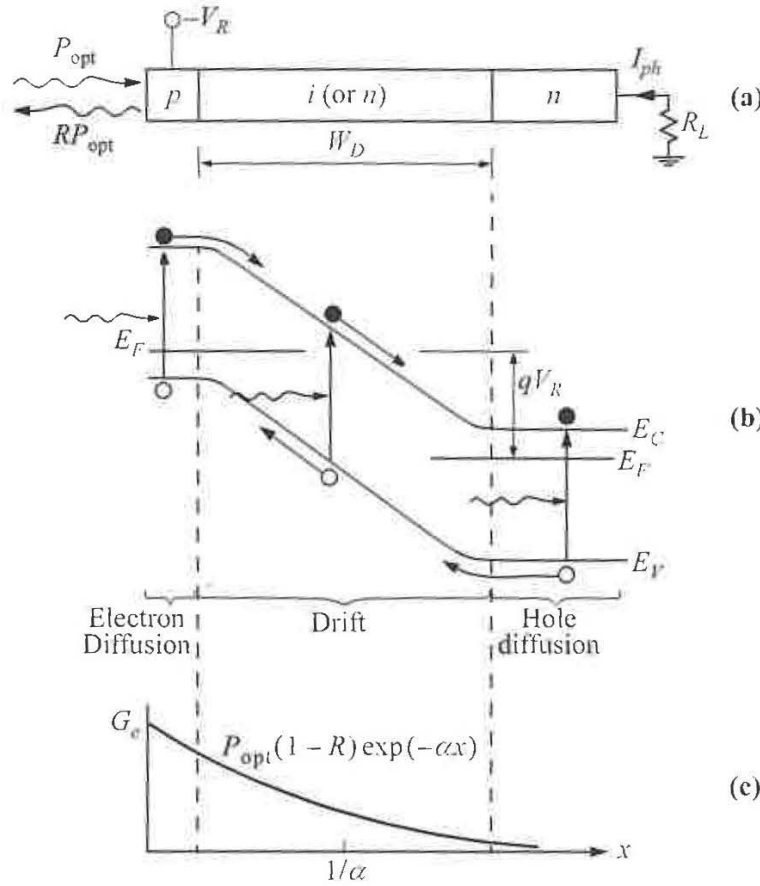


Fig. 6 Operation of photodiode. (a) Cross-sectional view of p - i - n diode. (b) Energy-band diagram under reverse bias. (c) Carrier generation characteristics. (After Ref. 1.)

$$J_{dr} = -q \int_0^{W_D} G_e(x) dx = q \Phi_0 [1 - \exp(-\alpha W_D)] \quad (27)$$

where W_D is the depletion-layer width. Note that within the depletion region, a quantum efficiency of 100% has been assumed.

For $x > W_D$, the minority-carrier density (holes) in the bulk semiconductor is determined by the one-dimensional diffusion equation

$$D_p \frac{\partial^2 p_n}{\partial x^2} - \frac{p_n - p_{no}}{\tau_p} + G_e(x) = 0 \quad (28)$$

where D_p is the diffusion coefficient for holes, τ_p the lifetime of excess carriers, and p_{no} the equilibrium hole density. The solution of Eq. 28 under the boundary conditions $p_n = p_{no}$ for $x = \infty$, and $p_n = 0$ for $x = W_D$ is given by

$$p_n = p_{no} - [p_{no} + C_1 \exp(-\alpha W_D)] \exp\left(-\frac{W_D - x}{L_p}\right) + C_1 \exp(-\alpha x) \quad (29)$$

with $L_p = \sqrt{D_p \tau_p}$ and

$$C_1 \equiv \left(\frac{\Phi_0}{D_p} \right) \frac{\alpha L_p^2}{1 - \alpha^2 L_p^2} \quad (30)$$

The diffusion current density is given by

$$\begin{aligned} J_{\text{diff}} &= -qD_p \left. \frac{\partial p_n}{\partial x} \right|_{x=W_D} \\ &= q\Phi_0 \frac{\alpha L_p}{1 + \alpha L_p} \exp(-\alpha W_D) + \frac{qp_{no}D_p}{L_p} \end{aligned} \quad (31)$$

The total current density is the sum of I_{dr} inside the depletion region and I_{diff} outside the depletion, given by

$$J_{\text{tot}} = q\Phi_0 \left[1 - \frac{\exp(-\alpha W_D)}{1 + \alpha L_p} \right] + \frac{qp_{no}D_p}{L_p} \quad (32)$$

Under normal operating conditions, the dark-current term involving p_{no} is much smaller so that the total photocurrent is proportional to the photon flux. The quantum efficiency can be obtained from Eqs. 2 and 32,

$$\eta = \frac{AJ_{\text{tot}}/q}{P_{\text{opt}}/h\nu} = (1 - R) \left[1 - \frac{\exp(-\alpha W_D)}{1 + \alpha L_p} \right] \quad (33)$$

Qualitatively, the quantum efficiency is reduced from unity due to reflection R and light absorbed outside the depletion region. For high quantum efficiency, low reflection R and $\alpha W_D \gg 1$ is desirable. However, for $W_D \gg 1/\alpha$, the transit-time delay may be considerable. We consider next the transit-time effect.

Frequency Response. Since the carriers require a finite time to traverse the depletion layer, a phase difference between the photon flux and the photocurrent will appear when the incident light intensity is modulated rapidly. To obtain a quantitative result for this effect, the simplest case is shown in Fig. 7a where all the light is assumed to be absorbed at the surface. The applied voltage is assumed to be high enough to deplete the intrinsic region and to ensure carrier saturation velocity v_s . For a photon flux density given by $\Phi_1 \exp(j\omega t)$ (photons/s-cm²), the conduction current density J_{cond} at point x is found to be, assume $\eta = 100\%$,

$$J_{\text{cond}}(x) = q\Phi_1 \exp \left[j\omega \left(t - \frac{x}{v_s} \right) \right] \quad (34)$$

The internal current is thus a function of time and distance. Since $\nabla \cdot J_{\text{tot}} = 0$, we can write the external total current as

$$J_{\text{tot}} = \frac{1}{W_D} \int_0^{W_D} \left(J_{\text{cond}} + \epsilon_s \frac{\partial \mathcal{E}}{\partial t} \right) dx \quad (35)$$

where the second term in parentheses is the displacement current. Substituting Eq. 34 into Eq. 35 yields

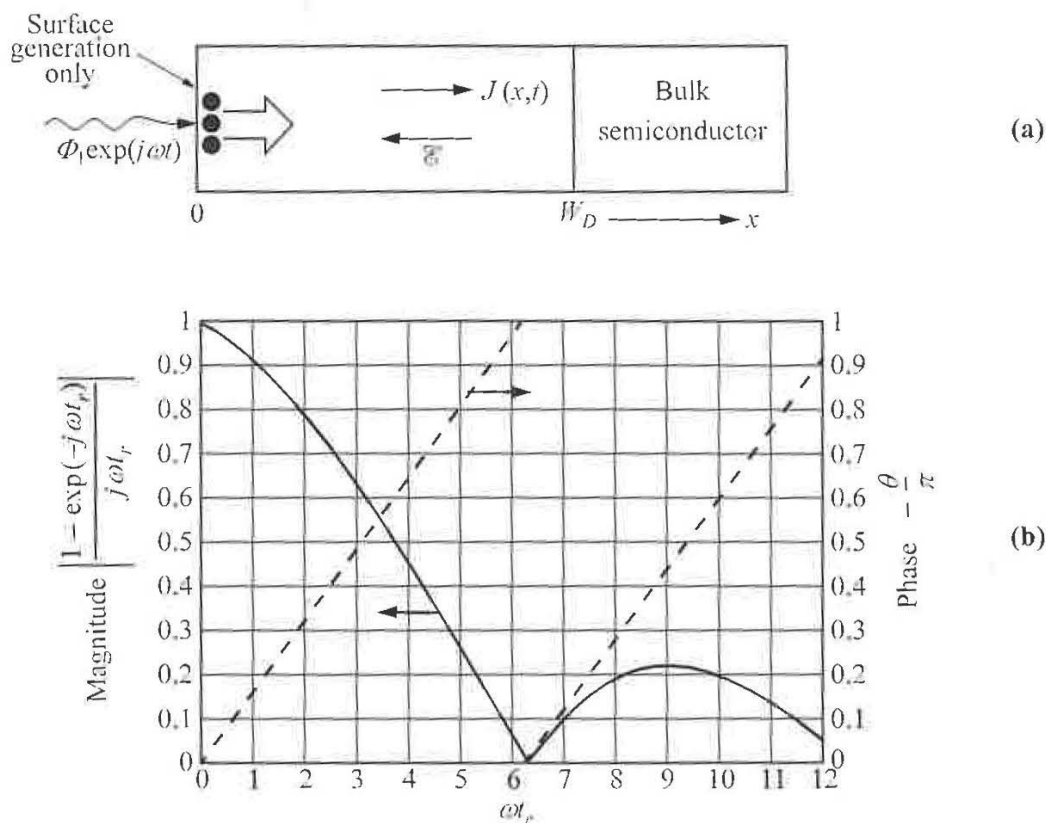


Fig. 7 (a) Geometry assumed for analysis of transit-time effect. Photoresponse (normalized magnitude and phase) vs. normalized modulation frequency of incident photon flux where $\theta = \omega t_r/2$. (After Ref. 10.)

$$J_{\text{tot}} = \left[\frac{j\omega \epsilon_s V}{W_D} + q\Phi_1 \frac{1 - \exp(-j\omega t_r)}{j\omega t_r} \right] \exp(j\omega t) \quad (36)$$

where V is the sum of applied voltage and the built-in potential, and $t_r = W_D/v_s$ is the transit time of carriers through the depletion region. From Eq. 36 the short-circuit current density ($V \approx 0$) is given by

$$J_{sc} = \frac{q\Phi_1 [1 - \exp(j\omega t_r)]}{j\omega t_r} \exp(j\omega t). \quad (37)$$

Figure 7b shows the transit-time effects at high frequencies where the amplitude and phase angle of the normalized current are plotted as a function of the normalized modulation frequency. Note that the magnitude of the ac photocurrent decreases rapidly with frequency when ωt_r exceeds unity. At $\omega t_r = 2.4$, the amplitude is reduced by $\sqrt{2}$ and is accompanied by a phase shift of 0.4π . The response time of the photodetector is thus limited by the carrier transit time through the depletion layer. A reasonable compromise between high-frequency response and high quantum efficiency is obtained for an absorption region of thickness $1/\alpha$ to $2/\alpha$ such that a reasonably large portion of the light is absorbed within the depletion region.

For the p - i - n photodiode, the thickness of the i -region is assumed equal to $1/\alpha$. The carrier transit time is the time required for carriers to drift through the i -region. From Eq. 37 the 3-dB frequency is given by ($\omega t_p = 2.4$)

$$f_{3\text{dB}} = \frac{2.4}{2\pi t_p} \approx \frac{0.4v_s}{W_D} \approx 0.4\alpha v_s \quad (38)$$

Figure 8 shows the internal quantum efficiency, that is, $\eta/(1 - R)$ of the Si p - i - n photodiode as a function of the 3-dB frequency and the depletion width calculated from Eq. 38 and Fig. 1. The curves illustrate the trade-off between the response speed (3-dB frequency which is proportional to $1/W_D$) and quantum efficiency at various wavelengths, by adjusting the depletion width.

The constructions of some high-speed photodiodes are shown in Fig. 9, usually with an antireflection coating (not shown) to increase quantum efficiency. The p - i - n photodiode is shown in Fig. 9a. The thickness of the intrinsic region (or low n -type doping, i -region, or low p -type doping, π -region) is optimized for the optical-signal wavelength and the modulation frequency. The p - n photodiode is a related structure where the n -type doping is high so that this layer is not fully depleted (Fig. 9b). At a wavelength close to the long-wavelength cutoff, the required absorption depth becomes very long (for $\alpha = 10 \text{ cm}^{-1}$, $1/\alpha = 1,000 \text{ }\mu\text{m}$). One option to compromise between quantum efficiency and response speed is to have the light incident from the side, parallel to the junction. This has the potential of reduced intrinsic layer thickness, shorter transit time, and thus higher speed, but at the expense of reduced quantum efficiency. The light can also be allowed to strike at an angle that creates multiple reflections inside the device, substantially increasing the effective absorp-

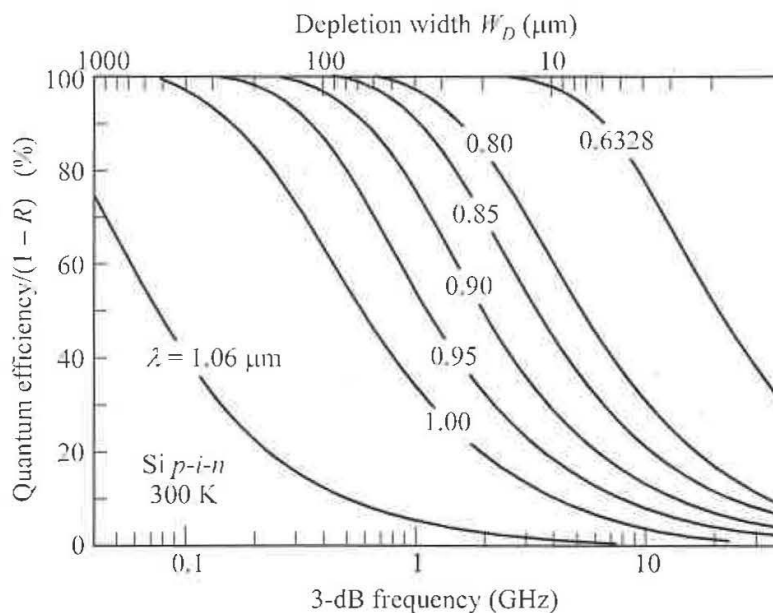


Fig. 8 Variation of quantum efficiency of Si p - i - n photodiode with depletion width and transit-time-limited 3-dB frequency for several wavelengths. Saturation velocity is 10^7 cm/s .

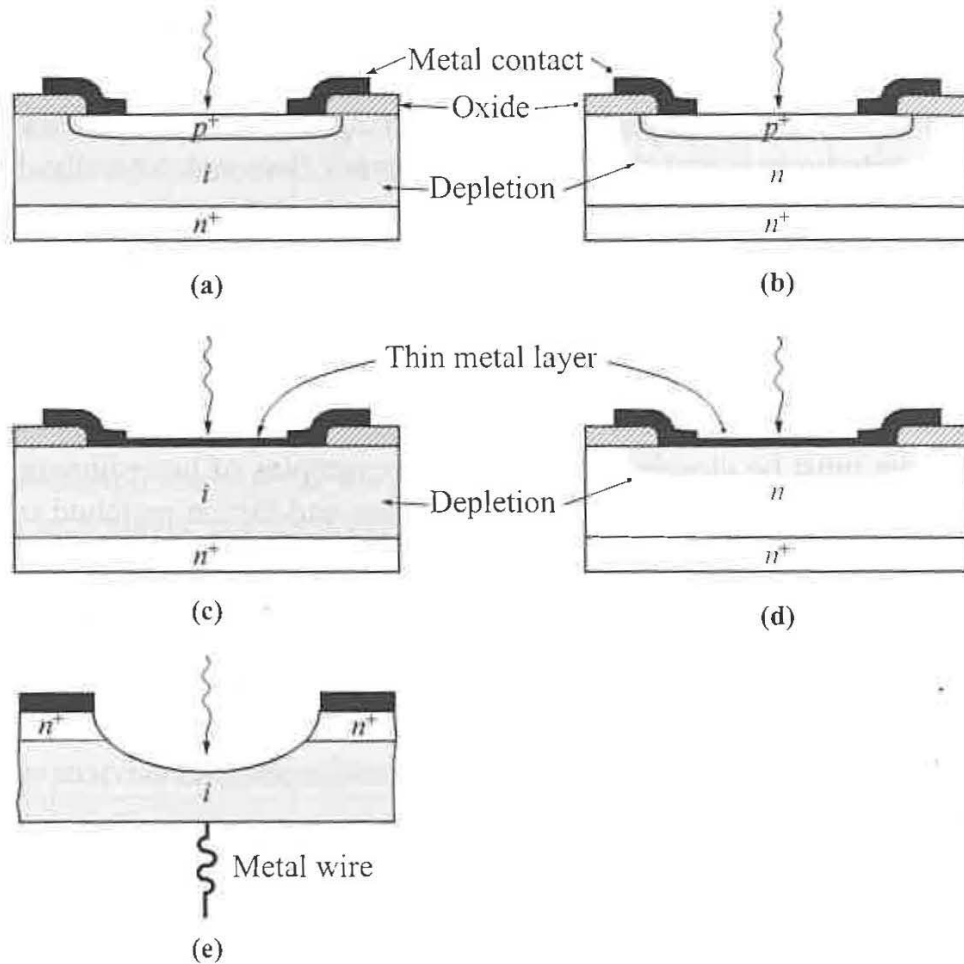


Fig. 9 Device configurations of some high-speed photodiodes. (a) *p-i-n* photodiode. (b) *p-n* photodiode. (c) Metal-*i-n* photodiode. (d) Metal-semiconductor photodiode. (e) Point-contact photodiode. (After Ref. 1.)

tion depth and at the same time keeping the carrier transit distance small.^{11,12} The other three devices are metal-semiconductor photodiodes, to be considered later.

For the *p-n* photodiode, since the depletion-layer is thin, some portion of light could be absorbed outside the depletion region. This can lead to some disadvantages. First, the quantum efficiency is reduced. Light absorbed outside the depletion region by more than a diffusion length does not contribute to photocurrent at all, and that within a diffusion length the efficiency is also reduced. Second, the diffusion process is a slow one. The time required for carriers to diffuse a distance x is known to be

$$t = \frac{4x^2}{\pi^2 D_p} \quad (39)$$

This is much slower than a drift process. The *p-n* photodiodes generally have a lower response speed than the *p-i-n* photodiodes. Finally, the neutral region contributes to series resistance which is a source of noise as discussed before.