Photonics and Lasers
An Introduction

Richard S. Quimby
Department of Physics
Worcester Polytechnic Institute
Worcester, MA

A Wiley-Interscience Publication
JOHN WILEY & SONS, INC.

Samsung Exhibit 1044, Page 1 of 34

Photonics and Lasers
An Introduction

Richard S. Quimby
Department of Physics
Worcester Polytechnic Institute
Worcester, MA

A Wiley-Interscience Publication
JOHN WILEY & SONS, INC.

Copyright © 2006 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-13 978-0-471-71974-8
ISBN-10 0-471-71974-9

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

Chapter 14

Photodiode Detectors

We saw in the previous chapter that photoconductive detectors suffer from two principle drawbacks: a poor time response, due to the long electron lifetime, and significant shot noise from the high level of dark current. Both of these problems are remedied in the *photodiode detector,* shown schematically in Fig. 14-1. In this device, light incident on the p–n junction of a semiconductor creates electron–hole pairs, which are swept out of the depletion region by the electric field there. Current flows in the external circuit only while charges are moving through this E field region, so the time response of the detector can be made quite fast. The p–n junction also provides a potential barrier for majority charge carriers, greatly reducing the amount of dark current and associated shot noise.

In this chapter, we discuss some important characteristics of photodiode detectors, including their behavior as electrical circuit elements, their time response, and their signal-to-noise properties.

14.1. BIASING THE PHOTODIODE

Although there are several types of circuits used to measure the photodiode signal current (see Section 14-5), the way that the photodiode is biased falls into one of two fundamental categories. In the *photovoltaic mode* (Fig. 14-2a), a load resistor R_L is directly connected across the photodiode, whereas in the *photoconductive mode* (Fig. 14-2b), the load resistor is connected through a series bias voltage V_B. In either case, the photocurrent generates a voltage V_R across the load resistor, which constitutes the detector output signal. The photoconductive mode we are discussing here should not be confused with the photoconductive-type detectors discussed in the previous chapter. The distinction is the presence or absence of a p–n junction in the device.

The current i in the circuit depends not only on the incident light intensity, but also on the values of R_L and V_B. To evaluate this current, we add up the potential changes around the circuit loop of Fig. 14-2b, and set the sum equal to zero (voltage loop law):

$$V_d + V_B + V_R = 0$$

We have adopted a sign convention for the diode voltage V_d and the resistor voltage V_R such that positive current flows into the positive side of each element. Of course, the actual values of the current or voltages may be either positive or negative. Writing $V_R = iR_L$ and solving for i, we have

$$i = \frac{-1}{R}(V_d + V_B) \qquad \text{(load line)} \tag{14-1}$$

Photonics and Lasers: An Introduction. By Richard S. Quimby

Copyright © 2006 John Wiley & Sons, Inc.

Samsung Exhibit 1044, Page 4 of 34

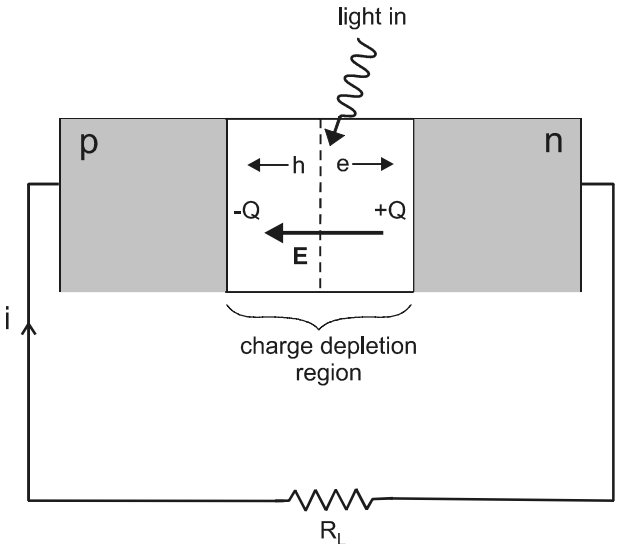

Figure 14-1 In a photodiode detector, the motion of electrons and holes across the charge depletion region causes a current in the external circuit. By convention, current *i* is defined as positive when it enters the p side of the diode. The photocurrent produced is therefore negative.

Equation (14-1) gives a relation between current *i* and diode voltage V_d that is imposed by the external circuit. Another relation between *i* and V_d comes from the internal constraints of the diode itself. The *i–V* relation for a semiconductor diode was given earlier in Eq. (10-21) and Fig. (10-14) for the case of no light absorption. When light is absorbed, the electron–hole pairs that are created cause an additional negative current, termed a *photocurrent.* The magnitude of the photocurrent is given by Eq. (13-7), which can be written here as

$$i_\lambda = \frac{P_{\text{in}}}{h\nu}\, e\eta_{\text{abs}} \qquad \text{(photocurrent)} \tag{14-2}$$

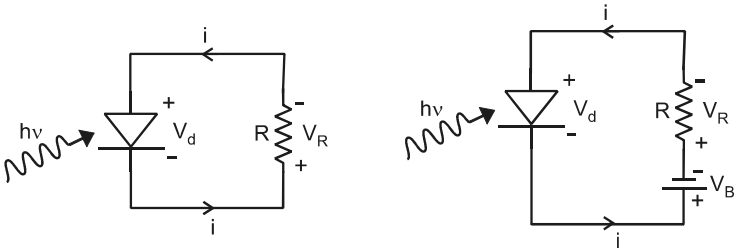

Figure 14-2 (a) In the photovoltaic mode, a load resistor is directly connected across the photodiode. (b) In the photoconductive mode, the load resistor in connected in series with a reverse-bias voltage.

where η_{abs} is the fraction of incident photons that are absorbed to create electron–hole pairs. This photocurrent adds to the diode current of Eq. (10-21) to give a total circuit current

$$i = i_0\left[\exp\left(\frac{eV_d}{\beta k_B T}\right) - 1\right] - i_\lambda \tag{14-3}$$

According to Eqs. (14-2) and (14-3), the diode *i–V* curve is shifted downward (along the $-i$ axis) by an amount i_λ, which is proportional to the incident light intensity. A few representative *i–V* curves for a photodiode are shown in Fig. 14-3, for equally spaced values of light intensity. For self-consistency, the diode current i and voltage V_d must satisfy both Eq. (14-1) and the diode *i–V* relation simultaneously. The solution for i can easily be obtained graphically, by plotting Eq. (14-1) on the same graph as the diode *i–V* curves. This procedure, shown in Fig. 14-3, is known as a *load-line analysis,* and Eq. (14-1) is known as the *load line.* A similar analysis was discussed in Section 11-1 in connection with biasing an LED.

Since the photovoltaic mode is just a special case of the photoconductive mode, with $V_B = 0$, both circuits can be analyzed in the same fashion using Fig. 14-3. The intersection of the load line and the diode *i–V* curve corresponds to the *operating point* of the circuit, which gives the value of both i and V_d. For the photovoltaic mode, the load line passes through the origin, so the operating point is always in the fourth quadrant, with positive V_d and negative i. In the photoconductive mode, the intercept on the V_d axis is at $V_d = -V_B$, so the operating point can be either in the third or fourth quandrants. The current i is always negative, but V_d can be either positive or negative.

The photovoltaic and photoconductive modes each have advantages and disadvantages, depending on the application. For low-level light detection, the photovoltaic mode has higher ultimate sensitivity than the photoconductive mode. This is because under dark conditions (no incident light), the photovoltaic operating point is at $i = 0$, whereas the photoconductive mode is at $i = -i_0$, the reverse saturation current. This minimum current

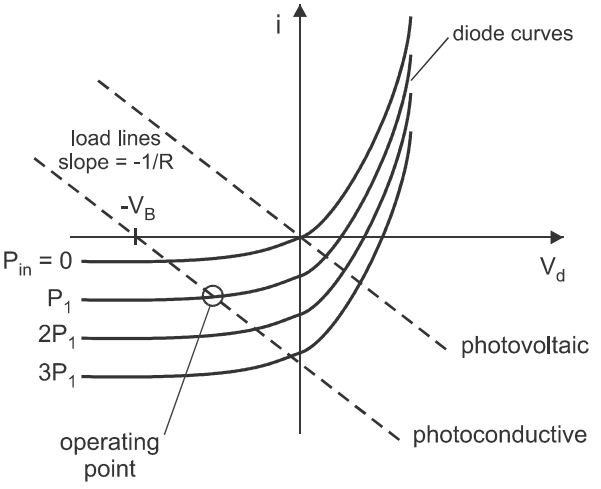

Figure 14-3 The operating point in a photodiode circuit is determined by the intersection between the load line and the diode *i–V* curve. Increasing optical power P_{in} shifts the *i–V* curve downward by an amount $i_\lambda \propto P_{in}$, moving the operating point down and to the right.

i_0 is termed the *dark current.* The shot noise from this dark current makes the photoconductive mode inherently more noisy. On the other hand, the photoconductive mode has a faster time response, and a linear response over a wider range of light intensities, as we shall see in the following sections.

One important application utilizing the photovoltaic mode is the *solar cell,* which converts optical power into electrical power. The electrical power supplied to the load resistor is $P_{\text{elec}} = i^2R_L$, where i is determined by Eq. (14-3) with $V_d = -iR$. Under practical conditions of solar illumination, $i_\lambda \gg i_0$, and Eq. (14-3) can be approximated as

$$i \simeq i_0 \exp\left(\frac{-eiR}{\beta k_B T}\right) - i_\lambda \qquad \text{(solar cell current)} \tag{14-4}$$

with i_λ given by Eq. (14-2). The efficiency

$$\eta_{sc} = \frac{P_{\text{elec}}}{P_{\text{in}}} = \frac{i^2R}{P_{\text{in}}} \qquad \text{(solar cell efficiency)} \tag{14-5}$$

of converting optical power into electrical power can then be calculated by solving Eq. (14-4) for i. Since this is an implicit equation for i, it must be solved numerically or graphically.

An important consideration for the solar cell is the choice of load resistance that maximizes the conversion efficiency. Since the load line in Fig. 14-3 has a slope $-1/R_L$, the operating point moves close to $i = 0$ for large R and $V_d = 0$ for small R. The power $|iV_d|$ delivered to the resistor will, therefore, have a maximum at some value of R. This optimum value of resistance can be determined graphically, or numerically as in the following example.

EXAMPLE 14-1

A silicon solar cell has an area of 4 cm^2, reverse saturation current density 1.5×10^{-8} A/cm^2, and diode ideality factor $\beta = 1$. Assume that light of intensity $I = 1000$ W/m^2 and average wavelength 500 nm is incident on the cell, and that 80% of the light is absorbed. Determine the optimum load resistance and power conversion efficiency. Repeat the calculation for $\beta = 2$.

Solution: The power striking the cell is (1000 W/m^2)(4×10^{-4} m^2) = 0.4 W. The photocurrent is then

$$i_\lambda = \frac{P_{\text{in}}\lambda}{hc}\, e\eta_{\text{abs}} = \frac{(0.4)(500 \times 10^{-9})(1.6 \times 10^{-19})(0.8)}{(6.63 \times 10^{-34})(3 \times 10^8)} = 0.129 \text{ A}$$

For room temperature (20°C = 293 K),

$$V_T \equiv \frac{k_B T}{e} = \frac{(1.38 \times 10^{-23})(293)}{1.6 \times 10^{-19}} = 0.0253 \text{ V}$$

is the "voltage equivalent of temperature." Putting this in Eq. (14-4) gives (for $\beta = 1$)

$$i = (6 \times 10^{-8}) \exp\left(\frac{-iR}{0.0253}\right) - 0.129$$

For a particular value of R, this equation is solved numerically for i, and the efficiency $\eta_{sc} = i^2R/P_{in}$ is calculated. By varying R, the graph shown in Fig. 14-4 is obtained. The optimum efficiency of 8.94% is obtained for $R = 2.5\Omega$.

If $\beta = 2$, Eq. (14-4) becomes

$$i = (6 \times 10^{-8}) \exp\left(\frac{-iR}{0.0506}\right) - 0.129$$

and the optimum efficiency increases to 17.9% at $R = 5\Omega$.

A more detailed model of solar cell efficiency would take into account the variation with wavelength of the optical power from the sun and the fraction of this light that is absorbed by the silicon. In practice, solar cells based on crystalline silicon can have efficiencies as high as 24% in the laboratory, with ~ 15% being typical in commercial devices. Thin films of amorphous silicon (atoms not ordered periodically) are inexpensive to manufacture but have lower efficiencies, typically 13% in the laboratory and 5–7% in commercial devices.

14-2. OUTPUT SATURATION

In the case of the solar cell just discussed, the primary goal is to convert as much optical power as possible into electrical power. When the photodiode is used as a light detector, however, it is generally more important that the detector output be linear with the incident light power. In this section, we examine the linearity of photodiode detector circuits using the two types of biasing modes.

Photovoltaic Mode

Consider first the photovoltaic bias mode shown in Fig. 14-2a. When R is very large (open-circuit condition), the load line is nearly horizontal, and the operating point is close to the V_d axis where $i \approx 0$. In that case, Eq. (14-3) becomes

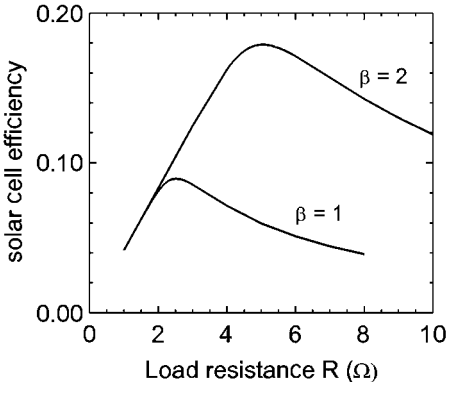

Figure 14-4 Variation of solar cell efficiency with load resistance for Example 14-1. Area of the cell is 4 cm^2. Optimum efficiency is higher and occurs at a higher load resistance when the diode ideality factor (β) is 2 rather than 1.

$$0 \simeq i_0\left[\exp\left(\frac{eV_d}{\beta k_B T}\right) - 1\right] - i_\lambda$$

with i_λ given by Eq. (14-2). Solving for the diode voltage gives

$$V_d = \frac{\beta k_B T}{e} \ln\left(1 + \frac{i_\lambda}{i_0}\right) \tag{14-6}$$

If the induced photocurrent is much greater than the dark current ($i_\lambda \gg i_0$), this becomes

$$V_d \simeq \beta V_T \ln\left(\frac{P_{\text{in}} e \eta_{\text{abs}}}{i_0 h\nu}\right) \qquad \text{(open circuit, } i_\lambda \gg i_0\text{)} \tag{14-7}$$

where

$$V_T \equiv \frac{k_B T}{e} \qquad \text{(voltage equivalent of temperature)} \tag{14-8}$$

The diode voltage, therefore, varies logarithmically with the incident power for $i_\lambda \gg i_0$. For $i_\lambda \ll i_0$, however,

$$V_d \simeq \beta V_T \frac{i_\lambda}{i_0} = \beta V_T \frac{P_{\text{in}} e \eta_{\text{abs}}}{i_0 h\nu} \tag{14-9}$$

Defining the quantity

$$R_{sh} = \frac{\beta V_T}{i_0} = \frac{\beta k_B T}{e i_0} \qquad \text{(shunt resistance)} \tag{14-10}$$

the diode voltage can be written as

$$V_d \simeq R_{sh} i_\lambda = R_{sh}\left(\frac{P_{\text{in}} e \eta_{\text{abs}}}{h\nu}\right) \qquad \text{(open circuit, } i_\lambda \ll i_0\text{)} \tag{14-11}$$

This result suggests that the photodiode can be modeled as an ideal current source connected in parallel with a resistor R_{sh}, as depicted in Fig. 14-5. Since $i = 0$ for an open cir-

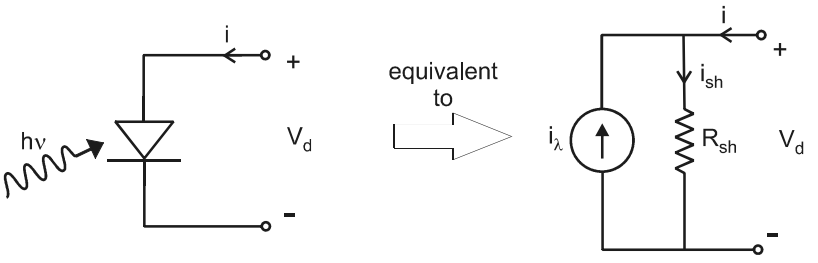

Figure 14-5 When $i_\lambda \ll i_0$, the photodiode can be modeled as an ideal current source in parallel with a shunt resistance R_{sh}.

cuit, the diode voltage becomes $V_d = i_\lambda R_{sh}$, in agreement with Eq. (14-11). Since R_{sh} appears in parallel with the current source, it is termed a *shunt resistance.* Higher values of R_{sh} are generally desirable, because the detector is then more sensitive to weak light signals (V_d large for small i_λ).

Values of shunt resistance vary widely, and are higher for wider bandgap materials, for which i_0 is smaller. R_{sh} is also higher at lower temperature (less thermal generation of electron–hole pairs) and for smaller junction area (since $i_0 = J_0 A$). For a typical room-temperature silicon photodiode with 1 cm^2 area, $R_{sh} \approx 10$ MΩ.

The response of the open-circuit photodiode to varying optical powers can be summarized as follows. At low incident power levels the response is linear with power, whereas at high power levels the response becomes logarithmic. The transition between these two regimes corresponds to $i_\lambda \sim i_0$, which is equivalent to $i_\lambda R_{sh} \sim V_T$. This deviation from linearity at high optical powers is referred to as *saturation* of the output signal, and is generally to be avoided.

To increase the range of optical powers over which the photodiode response is linear, the load resistance R_L can be made small. This makes the load line in Fig. 14-3 nearly vertical, intersecting the diode curves close to the current axis ($V_d \approx 0$). Since the diode curves are approximately evenly spaced for $V_d \leq 0$, the operating point moves downward along the $-i$ axis in proportion to the optical power. The voltage across the resistor, $V_R = iR_L$, is therefore linear with the optical power, as desired.

This conclusion can also be arrived at analytically. If $V_d \ll V_T$, the approximation $e^x \approx 1 + x$ allows Eq. (14-3) to be written as

$$i \simeq i_0\left(\frac{V_d}{\beta V_T}\right) - i_\lambda$$

or

$$i \simeq \frac{V_d}{R_{sh}} - i_\lambda \tag{14-12}$$

Using $V_d = -iR_L$ for the photovoltaic mode (Fig. 14-2a) and solving for i gives

$$i \simeq \frac{-i_\lambda}{1 + R_L/R_{sh}} \tag{14-13}$$

Since $i_\lambda \propto P_{\text{in}}$, we conclude that $V_R = iR_L \propto P_{\text{in}}$, as desired.

The equations above can be understood in terms of the equivalent circuit shown in Fig. 14-6. As before, the photodiode is represented as an ideal current source shunted by the resistance R_{sh}. The load resistor R_L is now connected in parallel with both of these. Defining the current i_{sh} through the shunt resistance to be positive in the downward direction, we have by the junction rule

$$i + i_\lambda = i_{sh}$$

This is equivalent to Eq. (14-12), using $i_{sh} = V_d/R_{sh}$. It is left as an exercise to show that Eq. (14-13) can be derived from this equivalent circuit model.

When $R_L \ll R_{sh}$, Eq. (14-13) becomes $i \simeq -i_\lambda$, and the diode voltage is

$$V_d = -iR_L \simeq i_\lambda R_L$$

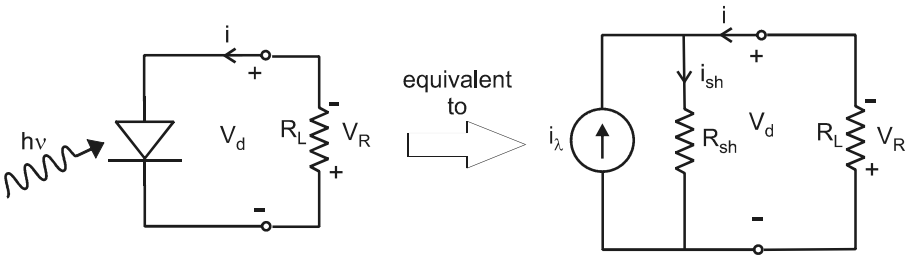

Figure 14-6 Equivalent circuit for a photodiode biased with load resistor R_L. This model is valid when $V_d \ll V_T$.

Defining the output of the detector as $V_{\text{out}} \equiv V_d$, we then have

$$V_{\text{out}} = V_d \simeq \frac{P_{\text{in}}}{h\nu} e\eta_{\text{abs}} R_L \qquad (V_d \ll V_T, R_L \ll R_{sh}) \tag{14-14}$$

where Eq. (14-2) has been used.

This result shows that under the two specified conditions the detector voltage V_d is linear not only with the incident optical power but also with the load resistance. In practice, it is easier to measure a larger voltage, so a larger R_L is desirable. However, as R_L is increased, one of these two conditions will eventually break down, and V_d will no longer increase with R_L. One possibility is that $R_L \ll R_{sh}$ breaks down, while the condition $V_d \ll V_T$ still holds. In the limit where $R_L \gg R_{sh}$, Eq. (14-13) gives

$$V_d = -iR_L \simeq i_\lambda R_{sh}$$

which becomes independent of R_L. This is the same result obtained in Eq. (14-11) for the open-circuit condition $R_L \to \infty$. Although the detector output V_{out} is no longer linear with R_L in this case, it is still linear with P_{in}.

The other possibility as R_L increases is that the condition $V_d \ll V_T$ breaks down first, in which case the exact expression in Eq. (14-3) must be used. Under these conditions, the output V_{out} is no longer linear with either R_L or P_{in}. According to Eq. (14-14), this saturation will occur at a certain value of the product $P_{\text{in}}R_L$. For higher R_L, saturation occurs at a lower P_{in}, and at higher P_{in}, saturation occurs at a lower R_L. There is, therefore, a trade-off between sensitivity (large output for small input) and dynamic range (range of inputs for which response is linear). The saturation with incident power for different values of load resistance is illustrated in Fig. 14-7.

Photoconductive Mode

Saturation behavior in the photoconductive mode can be understood by referring to the load-line analysis of Fig. 14-3. The load line has a slope $-1/R_L$, with an intercept on the voltage axis of $V_d = -V_B$. As the incident optical power increases, the operating point moves downward and to the right along the load line, decreasing the magnitude of reverse-bias voltage and increasing the magnitude of the current. Both the voltage and current change linearly with increasing optical power, until the operating point reaches the

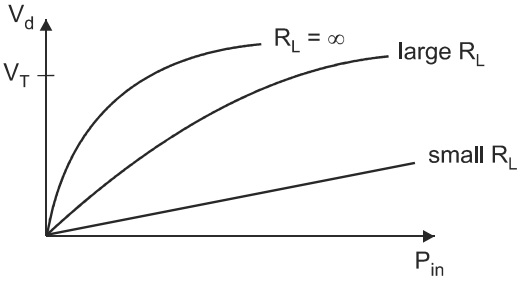

Figure 14-7 Diode voltage versus incident optical power for the photovoltaic mode. Smaller load resistance R_L gives a larger dynamic range but lower sensitivity. For $R_L \to \infty$, the effective resistance reaches the upper limit of R_{sh}.

current axis ($V_d = 0$). At that point, the detector circuit saturates, and the output is no longer linear with the incident optical power.

In the linear regime, we can obtain a simple analytical expression for the detector signal as follows. It is clear from Fig. 14-3 that $V_d < 0$ in the linear regime. Eq. (14-3) can then be written as

$$i = i_0\left[\exp\left(\frac{-|V_d|}{\beta V_T}\right) - 1\right] - i_\lambda$$

Unless the operating point is close to saturation, it is a good approximation that $|V_d| \gg V_T$. The exponential term above can then be neglected, giving

$$i \simeq -i_0 - i_\lambda \tag{14-15}$$

The detector output in the photoconductive mode is generally taken to be the voltage V_R across the load resistor. Since this will always be negative, we define the output as $V_{\text{out}} \equiv -V_R$ to give a positive number. Therefore,

$$V_{\text{out}} = -V_R = -iR_L \simeq (i_0 + i_\lambda)R_L \tag{14-16}$$

which can be written as

$$V_{\text{out}} \simeq i_0 R_L + \frac{P_{\text{in}}}{h\nu} e\eta_{\text{abs}} R_L \qquad (V_{\text{out}} < V_B) \tag{14-17}$$

using Eq. (14-2).

The detector output is seen to have two components, one proportional to the incident optical power, and the other independent of power. The component that varies with P_{in} is identical to the expression obtained in Eq. (14-14) for the photovoltaic mode. In both cases, the output voltage arises from the photocurrent i_λ flowing through load resistor R_L. The detector output can be expressed more compactly by defining the *responsivity* of the detector as

$$\mathcal{R} \equiv \frac{i_\lambda}{P_{\text{in}}} = \frac{e\eta_{\text{abs}}}{h\nu} \tag{14-18}$$

which is similar to the definition given in Eq. (13-8) for emissive-type photodetectors. The output in the photoconductive mode is then

$$V_{\text{out}} = (i_0 + \mathcal{R}P_{\text{in}})R_L \tag{14-19}$$

A similar expression applies to the photovoltaic mode, but without the i_0 term.

According to Eq. (14-17), the change in output voltage is linear with both R_L and P_{in}. However, this relation will only hold as long as $V_d < 0$, which requires that $V_{\text{out}} < V_B$. If P_{in} is increased above the point where $V_{\text{out}} \approx V_B$, the output saturates, and becomes approximately independent of P_{in}. This behavior is illustrated in Fig. 14-8 for two values of R_L. Larger R_L makes the detector more sensitive, since there is a larger output for a small value of P_{in}. However, this reduces the range of P_{in} over which the response is linear. The result is a sensitivity-dynamic range trade-off similar to that of the photovoltaic mode.

Although the photoconductive and photovoltaic modes have the similarities mentioned above, there are some significant differences. One difference is that saturation occurs at $V_{\text{out}} \approx V_B$ in the photoconductive mode, but at only $V_{\text{out}} \approx V_T$ in the photovoltaic mode. Since $V_T \approx 0.025$ V at room temperature, whereas V_B is typically several volts, the detector output in the photoconductive mode can be approximately two orders of magnitude larger than in the photovoltaic mode. This means that for the same detector sensitivity (same R_L), the dynamic range is approximately two orders of magnitude larger in the photoconductive mode than in the photovoltaic mode. This improved dynamic range is an important advantage of the photoconductive mode.

Another difference is that the photoconductive mode has a dark current, whereas the photovoltaic mode does not. The presence of dark current has two consequences. First, it contributes a constant background level that must be subtracted from the detector output in order to obtain the "true" signal (the signal arising from the incident light). Second, it contributes shot noise to the detector output, as discussed in Section 13-3. If the optical power is sufficiently large so that $i_\lambda \gg i_0$, then both of these effects become unimportant. In this large-signal regime, the photoconductive mode is the best choice for the detector circuit. However, if $i_\lambda \ll i_0$, then shot noise from the dark current can become a dominant source of detector noise. In this small-signal regime, the photovoltaic mode is a better choice, in order to obtain the best possible signal-to-noise ratio. The signal-to-noise properties of detector circuits are further discussed in Section 14-5.

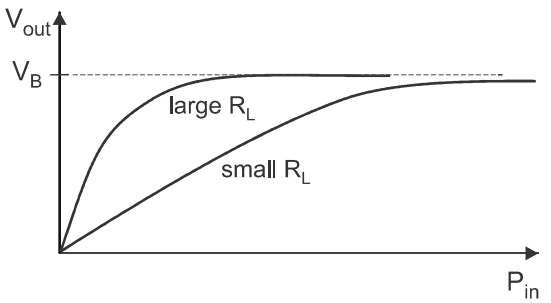

Figure 14-8 For a photodiode biased in the photoconductive mode, the detector response is linear for output voltages up to the reverse-bias voltage V_B. Larger load resistance R_L gives higher sensitivity but smaller dynamic range.

It should be emphasized that the dark current in a reverse-biased photodiode detector is much smaller and more well-defined than that in a photoconductive-type detector (one without a p–n junction). For example, a 1 cm^2 silicon photodiode has a typical dark current $i_0 \approx 1.5 \times 10^{-8}$ A, independent of reverse-bias voltage. In contrast, a CdS photocell has a background current that depends on the applied voltage, a typical value being ~ 10^{-5} A for a similar cross-sectional area and applied bias of 10 V.

14-3. RESPONSE TIME

An important characteristic of any photodetector is its response time—the time it takes for the detector output to change in response to changes in the input light intensity. It was noted in Chapter 13 that the response time of photoconductive-type detectors is quite poor because of the electron replenishment process, which keeps the induced photocurrent flowing for the duration of the electron's lifetime. In photodiode detectors, this replenishment process is suppressed by the p–n junction, which presents a barrier to the movement of majority carriers. The response time is thereby significantly improved, since the time taken for charge carriers to move through the high-field region of the junction (the carrier *transit time*) can be much shorter than the carrier lifetime. This improved time response is countered in part, however, by the capacitance of the p–n junction and associated RC time constant. In this section, we consider the implications and relative importance of transit time and capacitance in determining the photodiode response time.

Junction Capacitance

The capacitance of a p–n junction can be evaluated by determining how the charge on either side of the junction changes in response to a changing diode voltage. We will assume for simplicity a highly doped p region and weakly doped n region, so $N_A \gg N_D$ as in Fig. 10-11. In this case, most of the charge depletion region is on the n side ($d \approx d_n \gg d_p$), and the junction width d is given by Eq. (10-20). When an external voltage V is applied to the diode, the internal potential V_0 is replaced by $V_0 - V$, giving

$$d = \sqrt{\frac{2\varepsilon(V_0 - V)}{eN_D}} \qquad \text{(p–n junction width)} \tag{14-20}$$

where V is positive for forward bias and negative for reverse bias. A change ΔV in external voltage causes a change in junction width of

$$\Delta d = \frac{1}{2d}\left(\frac{-2\varepsilon}{eN_D}\right)\Delta V \tag{14-21}$$

as can be verified by taking the differential of both sides of Eq. (14-20). Most of this change in width occurs on the n side of the junction, since $d_n \gg d_p$. Writing the charge of the uncovered ion cores in the n region as $Q = eN_D(Ad_n)$, the change in this charge is

$$\Delta Q = eN_D(A\Delta d) \tag{14-22}$$

There is an equal but opposite change in the charge on the p side of the junction. Combining Eqs. (14-21) and (14-22), the junction capacitance is then

$$C \equiv \left| \frac{\Delta Q}{\Delta V} \right| = \frac{\varepsilon A}{d} \tag{14-23}$$

which is the familiar expression for the capacitance of a parallel-plate capacitor of plate area A and spacing d, separated by a medium with dielectric constant ε. Using Eq. (14-20) for d gives the result

$$C = A\sqrt{\frac{eN_D\varepsilon}{2(V_0 - V)}} \qquad \text{(p–n junction capacitance)} \tag{14-24}$$

For reverse-bias voltage $V = -V_B$, this becomes

$$C = A\sqrt{\frac{eN_D\varepsilon}{2(V_0 + V_B)}} \tag{14-25}$$

It is seen that the capacitance of the p–n junction is not constant, but rather decreases with increasing V_B. This is a consequence of the junction width d increasing with V_B.

The junction capacitance can be considered to be in parallel with the diodes in Fig. 14-2, which leads to a first-order RC circuit time response. If the incident power is suddenly switched from zero to some constant value at $t = 0$, the output voltage increases exponentially in time according to

$$V_{\text{out}}(t) = V_{\max}(1 - e^{-t/RC}) \tag{14-26}$$

which is shown graphically in Fig. 14-9. The product RC is known as the *time constant* of the circuit, and has units of seconds with R in ohms and C in farads. The time constant measures how quickly the output responds to a changing input. For example, at time $t = RC$ the output has risen to 63.2% of the final steady-state value, and at time $t = 2RC$ it has risen to 86.5% of this value.

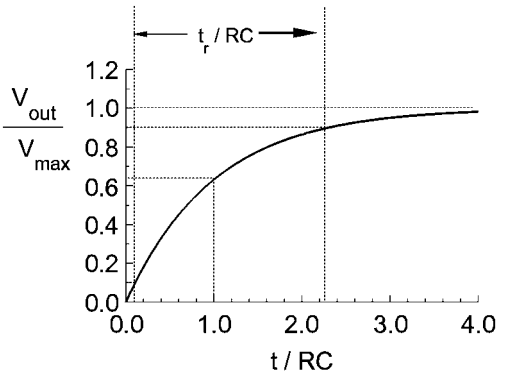

Figure 14-9 In an exponential rise, the time constant RC gives the time taken for the output to reach 63.2% of the final steady-state value. The rise time t_r is the time taken for the output to go from 10% to 90% of the final value.

An alternative measure of the time response is the *rise time,* defined as the time taken to rise from 10% to 90% of the final value. For the RC circuit, it is straightforward to show (see Problem 14.6) that

$$t_r = (\ln 9)RC = 2.2\ RC \qquad \text{(rise time for RC circuit)} \tag{14-27}$$

The rise time is a more general definition for time response than the time constant, because it applies equally well to an exponential or nonexponential time dependence. When the time response is nonexponential, Eq. (14-27) can be interpreted as defining an effective time constant in terms of the rise time.

The rise time or RC time constant characterizes the detector time response to a step-function intensity modulation, in which the incident light intensity is suddenly changed from one value to another. If instead the incident light intensity is sinusoidally modulated, then the output will be sinusoidally modulated as well. These two different types of modulation were discussed in connection with the LED time response (see Fig. 11-3). The output amplitude is approximately independent of frequency up to a limiting upper value, the bandwidth, above which the amplitude becomes smaller. Denoting the 3 dB electrical bandwidth as B, we have

$$B = \frac{1}{2\pi RC} \tag{14-28}$$

where Eq. (11-13) has been used with $\tau = RC$ and $B = f_e$. The bandwidth can be written in terms of the rise time as

$$B = \frac{2.2}{2\pi t_r} = \frac{0.35}{t_r} \tag{14-29}$$

using Eq. (14-27). This last expression can be taken as defining the 3 dB bandwidth in the case of nonexponential time response.

The above results show that a smaller capacitance leads to a faster time response and larger bandwidth, which is usually desirable for a photodetector. According to Eq. (14-25), there are several parameters that can be adjusted to reduce the capacitance. For example, the reverse-bias voltage V_B can be increased. This makes the photoconductive mode ($V_B > 0$) inherently faster than the photovoltaic mode ($V_B = 0$). Indeed, this is another characteristic advantage of the photoconductive mode, in addition to the increased dynamic range that was discussed earlier. There is a practical limit to V_B, however, due to electrical breakdown in the junction. Typical reverse-bias voltages are 5–10 V.

Another way to reduce the capacitance is to decrease the junction area A. Smaller detector areas, therefore, give a faster time response. The downside of this approach is that it may not be possible to direct all the available light onto the semiconductor material if A is too small. The ability to focus light onto a small detector area is governed by the brightness of the light source, as discussed in Appendix A (see also Chapter 15). High-brightness sources such as a laser can be focused to a very small area ($\sim \lambda^2$), which allows a small detector area to be used without loss of efficiency. Low-brightness sources such as an LED or incandenscent filament, however, can be imaged only onto a much larger area. If a photodiode with small A is used to detect light of low brightness, much of the light to

be measured will not strike the detector's active area, making the detector less sensitive. This results in a sensitivity–speed trade-off, which must be optimized for best performance in a particular application.

The other parameter in Eq. (14-25) that can be adjusted to reduce the capacitance is N_D, the density of donors on the weakly doped n side of the junction. The capacitance is reduced when N_D is made smaller, because the junction width d then increases. This is one reason that in most photodiodes, one side of the junction is very weakly doped. There are other reasons for this also, as we will soon see.

Reducing the capacitance is one way to reduce the RC time constant, but it is not the only way. Reducing the load resistance R_L has the same effect, although this decreases the detector sensitivity, as seen in Eq. (14-19) and Fig. 14-8. Large R_L is best for high sensitivity, and small R_L is best for high speed, resulting in another sensitivity–speed trade-off. The dynamic range is also better for small R_L, as previously discussed. These various trade-offs are summarized in Fig. 14-10.

Carrier Transit Time

Single p–n Junction

When the RC time constant is made sufficiently small, the photodiode response time will be limited by the motion of charge carriers across the device. The fundamental principle needed for this analysis is given by Eq. (13-9), developed in connection with the vacuum photodiode. According to this result, which applies quite generally to any photon-type detector, the current pulse from a single photoexcited electron lasts as long as the electron is moving through a region with high electric field. A similar relation applies to photoexcited holes. Since the E field is high only in the depletion region (see Fig. 10-11), the current pulse will last a time

$$t_{tr} = \frac{d}{v} \tag{14-30}$$

known as the *transit time,* where d is the width of the depletion region and v is the velocity of the charge carrier. At low to moderate electric field strength, the electron velocity is $v_e = \mu_e E$, where μ_e is the electron *mobility,* and the hole velocity is $v_h = \mu_h E$. The time re-

	sensitivity	dynamic range	time response
large R	high	low	slow
small R	low	high	fast
large area	high	—	slow
small area	low	—	fast

Figure 14-10 Summary of performance trade-offs in choosing load resistor R_L and detector area A.

sponse will be limited by the charge carrier with the smallest mobility, which is usually a hole. In this case, the transit time becomes

$$t_{tr} = \frac{d}{\mu_h E} \qquad \text{(transit time, low field)} \tag{14-31}$$

If the field is sufficiently high (above a value $E_{sat} \sim 2 \times 10^6$ V/m for holes in Si), then the carrier velocity is no longer proportional to E, saturating at the upper limit v_s ($\sim 10^5$ m/s for holes in Si). Under high-field conditions, the transit time is

$$t_{tr} = \frac{d}{v_s} \qquad \text{(transit time, high field)} \tag{14-32}$$

It would appear from Eq. (14-31) that a higher E field gives a shorter transit time. However, the width of the depletion region d increases with increasing E, and this tends to increase the transit time. To see how these two effects offset each other, we express both E and d in terms of the junction potential $V_0 + V_B$. Taking E as approximately constant across the depletion region,

$$E \simeq \frac{V_0 + V_B}{d}$$

the transit time then becomes

$$t_{tr} = \frac{d^2}{\mu_h (V_0 + V_B)} = \frac{2\varepsilon}{\mu_h e N_D} \qquad \text{(transit time, low field)} \tag{14-33}$$

where Eq. (14-20) has been used with $V = -V_B$.

This result shows that the transit time with low field is actually independent of the applied reverse bias. Apart from the choice of semiconductor material, which determines ε and μ_h, it depends only on the donor concentration N_D. Higher N_D would appear to be best for a fast time response (small t_{tr}). However, we found in Eq. (14-25) that the junction capacitance increases with higher N_D, resulting in a slower response. The time response will, therefore, be optimized when the contributions of capacitance and transit time to the response time are approximately equal. The value of N_D giving this optimum response depends on several other parameters, as shown in the following example.

EXAMPLE 14-2

A p^+–n silicon photodiode has an junction area of 1 mm^2, and a reverse bias of 10 V is applied through a 10 kΩ load resistor. Determine the doping level in the lightly doped n region that minimizes the response time, and determine the junction width for this doping level. Take the hole mobility in Si to be 5×10^{-2} m^2/Vs.

Solution: The optimum time response will occur when the capacitance rise time 2.2 $R_L C$ is approximately equal to transit time t_{tr}. Using Eqs. (14-25) and (14-33),

$$2.2 R_L A \sqrt{\frac{e\varepsilon_0 \varepsilon_r N_D}{2V_B}} = \frac{2\varepsilon_0 \varepsilon_r}{\mu_h e N_D}$$

where $\varepsilon_0 = 8.85 \times 10^{-12}$ F/m is the permitivity of free space, $\varepsilon_r = 11.9$ is the relative dielectric constant for silicon, and $V_0 + V_B \approx V_B = 10$ V. Solving this for N_D gives

$$N_D = \frac{1}{e}\left[\frac{1.65\ \varepsilon_0\varepsilon_r V_B}{R_L^2 A^2 \mu_h^2}\right]^{1/3} = \frac{1}{1.6 \times 10^{-19}}\left[\frac{1.65(8.85 \times 10^{-12})(11.9)(10)}{(10^4)^2(10^{-6})^2(5 \times 10^{-2})^2}\right]^{1/3}$$

$$N_D = 1.18 \times 10^{18}\ \text{m}^{-3} = 1.18 \times 10^{12}\ \text{cm}^{-3}$$

This is a very light doping level, only two orders of magnitude above the "intrinsic" carrier concentration in undoped silicon of $\sim 1.4 \times 10^{10}$ cm^{-3}. The junction width is then obtained from Eq. (14-20) as

$$d \simeq \sqrt{\frac{2(8.85 \times 10^{-12})(11.9)(10)}{(1.6 \times 10^{-19})(1.18 \times 10^{18})}} = 1.06 \times 10^{-4}\ \text{m} = 0.106\ \text{mm}$$

The electric field in the depletion region is $E \simeq 10\ \text{V}/10^{-4}\ \text{m} = 10^5$ V/m. This is less than the saturating field value $E_{\text{sat}} \sim 2 \times 10^6$ V/m, which justifies using Eq. (14-33) for the transit time.

14-4. TYPES OF PHOTODIODES

PIN Photodiode

The analysis of the transit time response in the previous section assumed that the photoexcited electron–hole pairs are created in the depletion region, where there is a strong E field. This is not always the case, however, as illustrated in Fig. 14-11a. Some light that is incident on the highly doped p^+ side passes completely through the depletion region, and is absorbed in the n-type region on the other side. In this latter region, the E field is very small, because the high free-carrier concentration makes the electrical conductivity high (and E is small inside a good conductor). According to Eq. (13-9), there is little contribution to the photocurrent when E is small. Therefore, the photoexcited charge carriers created outside the depletion region do not contribute significantly to the photocurrent, as long as they remain outside the high-field depletion region.

The charge carriers created outside the depletion region do not remain motionless, however. Like any particles subject to random thermal motion, they spread out from their initial position in a process known as *diffusion.* Diffusion proceeds much more slowly than *drift,* which is the term given to the directed motion induced by an electric field. After spreading out by diffusion for a certain time, some holes initially generated in the n-type region will enter the depletion region, where they are quickly swept across by the high E field there. This results in an additional component to the photocurrent, delayed by the diffusion time. Holes generated at different distances from the edge of the depletion region have different diffusion times, which results in a diffusion "tail" in the photocurrent response to a square-wave light pulse. This type of signal distortion, depicted in Fig. 14-11b, is generally undesirable.

The solution to the problem of diffusing charge carriers is to simply eliminate the diffusion region. This can be accomplished by decreasing the donor concentration in the n

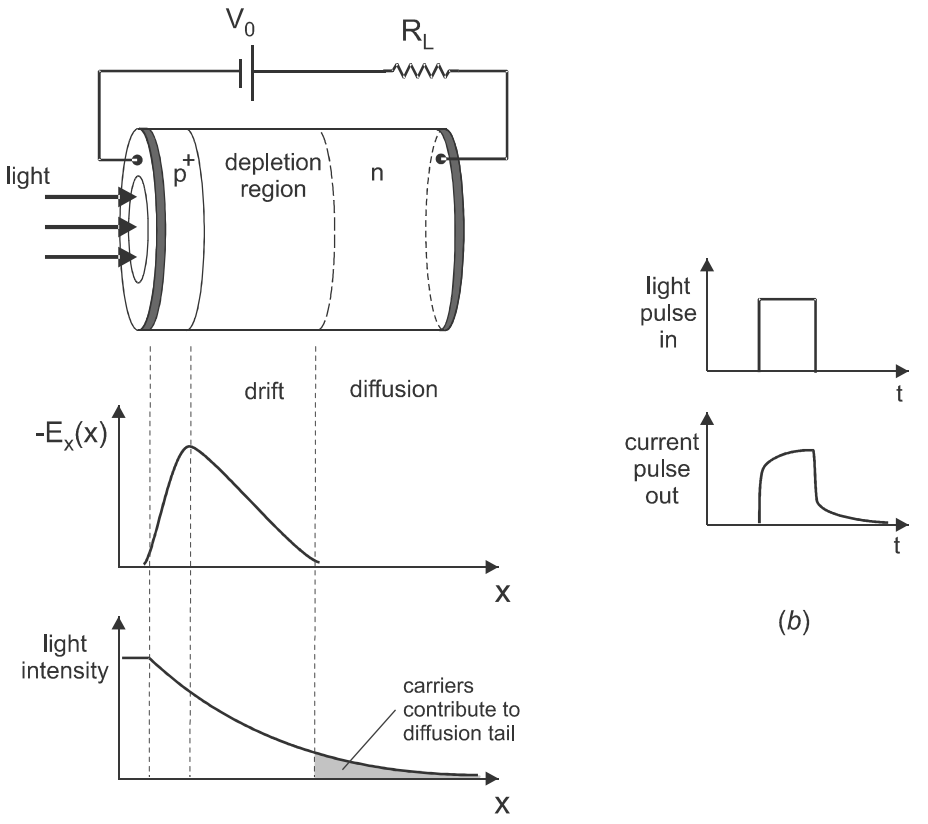

Figure 14-11 (a) In a simple p–n junction photodiode, charge carriers may be created in a high-field drift region or a low-field diffusion region. (b) Charge carriers created in the diffusion region give rise to a distortion in the photocurrent waveform.

region until the depletion region occupies nearly the entire space between the electrodes. As shown in Fig. 14-12, the *E* field then extends nearly all the way to the far electrode, so charge carriers generated anywhere in the material will be subject to drift rather than diffusion. At the far end is a thin, highly doped *n* region, needed to make good ohmic contact with the electrode. Since the middle region is very lightly doped (nearly intrinsic), it is labelled I, and the device is termed a *PIN photodiode.*

The PIN photodiode is the most commonly used photon detector today. It not only eliminates carrier diffusion, but also has the advantage that the depletion width *d* is fixed by the geometry of the device. The ability to adjust *d* by design, rather than applied voltage, allows the photodiode's performance to be optimized for specific applications. For example, making *d* larger increases the path length for absorption of light, which increases the efficiency η_{abs} with which light is absorbed. This is especially important for wavelengths near the semiconductor's bandgap, where $\alpha d \ll 1$. In this regime, Eq. (13-15) gives $\eta_{abs} \simeq (1 - R)\alpha d$.

On the other hand, a larger *d* degrades the time response by increasing the transit time. In Eqs. (14-31) and (14-32), *d* should now be considered a constant, independent of applied voltage. The transit time is therefore minimized by increasing *E* to the saturating value E_s.

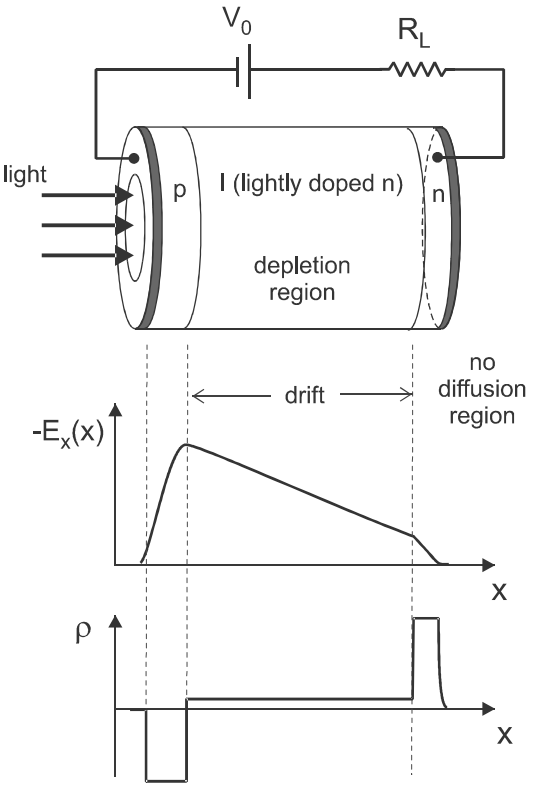

Figure 14-12 (a) In a PIN photodiode, charge carriers are mostly created in the high-field drift region, which extends almost to the far electrode. The lightly n-doped "intrinsic" region has a nearly constant *E* field, which sweeps charge carriers through the device without diffusion.

EXAMPLE 14-3

A silicon PIN photodiode has an intrinsic region of thickness 0.1 mm. Determine the minimum rise time for the detector, its corresponding bandwidth, and the required reverse-bias voltage. Repeat for an intrinsic region of thickness of 10 μm.

Solution: The minimum transit time for holes in silicon is

$$t_{tr} \simeq \frac{d}{v_s} = \frac{10^{-4}\ \text{m}}{10^5\ \text{m/s}} = 10^{-9}\ \text{s} = 1\ \text{ns}$$

Taking this as the rise time, the corresponding bandwidth is

$$B = \frac{0.35}{t_r} = \frac{0.35}{10^{-9}\ \text{s}} = 350\ \text{MHz}$$

The required *E* field is $E = E_s \simeq 2 \times 10^6$ V/m, so

$$\Delta V = Ed = (2 \times 10^6\ \text{V/m})(10^{-4}\ \text{m}) = 200\ \text{V}$$

which is an inconveniently high voltage. Repeating the calculation for $d = 10$ μm gives $t_{tr} = 0.1$ ns, $B = 3.5$ GHz, and $\Delta V = 20$ V. This is a much improved time response, and occurs with a more convenient bias voltage.

The above example shows that in terms of transit time, a thinner intrinsic region is preferable. However, if d is made too small, capacitive effects become important. There is, therefore, an optimum value of d that minimizes the overall response time (see Problem 14.8). Another problem with small d is illustrated by the following example.

EXAMPLE 14-4

For the silicon PIN photodiodes of Example 14-3, determine the absorption efficiency for 860 nm light. At this wavelength, the absorption coefficient in silicon is 335 cm^{-1} and the reflectivity (from air) is 32%.

Solution: For $d = 0.1$ mm,

$$\alpha d = (335\ \text{cm}^{-1})(10^{-2}\ \text{cm}) = 3.35$$

Eq. (13-15) then gives

$$\eta_{\text{abs}} = (1 - 0.32)(1 - e^{-3.35}) = 0.656$$

Note that this neglects light reflected back from the far end of the Si material, so it somewhat underestimates η_{abs}. Repeating the calculation for $d = 10$ μm gives $\alpha d = 0.335$, so

$$\eta_{\text{abs}} = (1 - 0.32)(1 - e^{-0.335}) = 0.194$$

The thinner intrinsic region is seen to be less efficient at absorbing the incident light. There is, therefore, a trade-off between detector speed and sensitivity. A PIN photodiode can be optimized for either of these, depending on the application.

Avalanche Photodiode

If a small load resistance R_L is used to increase the frequency bandwidth of a PIN photodiode, the signal voltage may be quite small, requiring amplification. This can be accomplished with electronic amplifiers, but these introduce their own sources of noise, and it is sometimes desirable to increase the signal generated by the detector, before amplification. One way to do this is through the avalanche multiplication process, depicted in Fig. 14-13a. This is the solid-state analog of the electron multiplication that takes place in a photomultiplier tube. An electron is accelerated by the E field in the depletion region of a reverse-biased p–n junction, and gains kinetic energy in proportion to the distance traveled. When the electron's kinetic energy is high enough, it can collide with an atom in the semiconductor and create an additional electron–hole pair, a process termed *impact ion-*

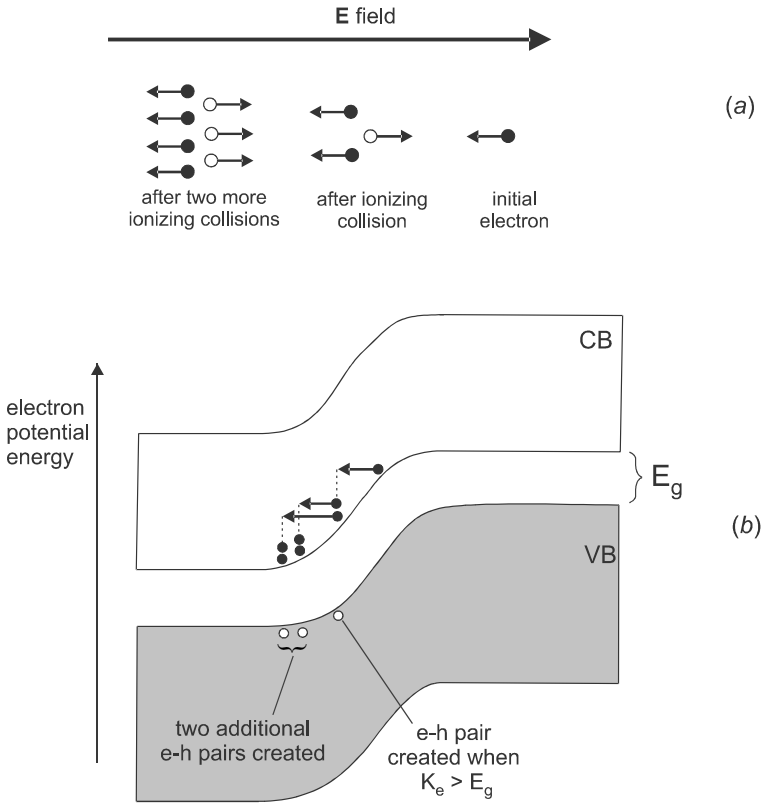

Figure 14-13 (a) An electron accelerated in a strong *E* field creates an additional electron–hole pair by impact ionization. Both the new electron and the original one then create additional electron–hole pairs, resulting in avalanche multiplication. (b) Impact ionization can occur when an electron in the CB picks up kinetic energy greater than the band-gap energy.

ization. There are now two electrons, and as each one of these accelerates in the *E* field, it can create yet another electron–hole pair by the same mechanism. There are now a total of four electrons, each of which can create another one to give eight, and so on. The result is *avalanche multiplication,* in which the number of charge carriers increases exponentially with distance traveled.

The creation of electron–hole pairs by impact ionization can be understood in terms of the energy band picture of Fig. 14-13b. After moving a distance Δx in a direction opposite to the *E* field, the electron loses an amount of potential energy of magnitude $eE\Delta x$. If no energy is lost to other processes, the electron then gains this same amount of kinetic energy *K*. When $K > E_g$, it becomes energetically possible for a collisional-energy transfer process to take place, in which the electron kinetic energy decreases by $\Delta K = -E_g$, while at the same time the potential energy of a valence electron is increased by this same amount. Increasing the potential energy of a valence electron by an amount E_g corresponds, in the band picture, to taking an electron out of the valence band and placing it in the conduction band, that is, creation of an electron–hole pair.

The energy needed to create an electron–hole pair is usually much less than the potential energy change of an electron as it moves across a reverse-biased p–n junction. For ex-

ample, a typical band-gap energy is $E_g \sim 1$–2 eV, whereas a 10 V reverse bias corresponds to a potential energy change of 10 eV. Since the electron acquires sufficient energy to create several electron–hole pairs as it moves through this potential difference, it might seem that avalanche multiplication should occur readily in most reverse-biased photodiodes. However, the avalanche phenomenon actually plays a minor role in conventional PIN photodiodes. The explanation for this is that the electron undergoes nonionizing collisions as well as ionizing collisions as it moves through the electric field. For example, the electron can scatter off the thermally induced vibrations in the material (lattice phonons), giving up kinetic energy to heat. At moderate electric field values, these nonionizing collisions prevent the electron's kinetic energy from reaching the threshold value $K = E_g$. When the electric field is sufficiently high, however, the electron can pick up kinetic energy $K > E_g$ before a nonionizing collision occurs, and the avalanche mechanism becomes more efficient.

The electric field at which the avalanche mechanism becomes important depends on the material, being higher in wider-band-gap materials that have a higher threshold kinetic energy. For silicon, an electric field $E \sim 5 \times 10^7$ V/m over a path length of ~ 2 μm is needed for efficient avalanche multiplication. This corresponds to a potential difference $\Delta V \sim 100$ V, much higher than the reverse bias of a typical PIN photodiode.

A photodiode utilizing avalanche multiplication to achieve gain is termed an *avalanche photodiode,* or APD. The structure of an APD, depicted in Fig. 14-14, differs from that of the PIN photodiode (Fig. 14-12) in two ways. First, light enters through a

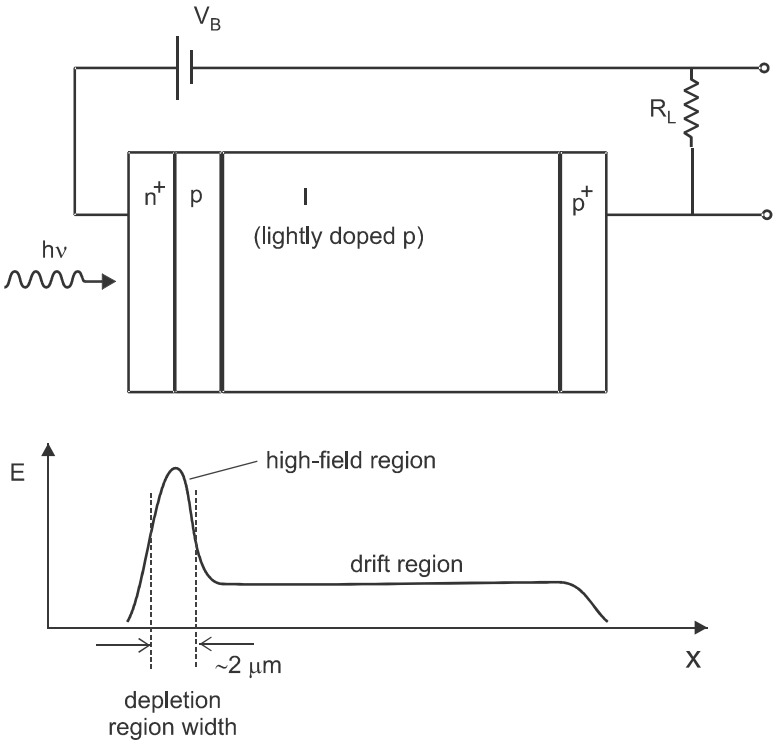

Figure 14-14 In an avalanche photodiode (APD), electrons photoexcited in a nearly intrinsic region are swept out by a small electric field there, and injected into a high-field region between highly doped n and p layers. Avalanche multiplication occurs primarily in the high-field region.

highly doped n-type layer rather than a p-type layer. Second, an additional p-type layer has been added between the highly doped n layer and the nearly intrinsic layer. The leftmost n and p layers are highly (and nearly equally) doped, so the junction width between them is small ($d \sim 2$ μm), and the high electric field is mostly confined to this region. In the adjacent intrinsic region (actually lightly doped p-type), there is a much smaller and nearly uniform E field, which extends out to the highly doped p^+ region on the right.

The APD operates in the following way. Light passes through the thin n^+ and p layers and is absorbed in the much thicker intrinsic region. Electrons and holes created by photoabsorption then drift in opposite directions under the influence of the E field, electrons to the left and holes to the right. The electrons eventually make it to the high-field region, where they undergo avalanche multiplication. Holes do not initiate the avalanche in this scheme, but those created by impact ionization can contribute to its development. In silicon, however, holes are much less efficient at causing ionization than are electrons, and therefore make only a minor contribution to the amplification.

In principle, an APD could be constructed with a p^+–n–i–n^+ structure, instead of the n^+–p–i–p^+ shown in Fig. 14-14. However, in this case it would be the holes that are injected into the high-field region, and in silicon this would result in weak amplification. For this reason the n^+–p–i–p^+ structure is always used for silicon APD's. This asymmetry between electron and hole ionization probabilities also has implications for the signal-to-noise properties of the APD. Since avalanche multiplication is a statistical process, there is less statistical variation in output current (i.e., less noise) when only one type of charge carrier contributes to the avalanche. For this reason, germanium APDs, in which the electrons and holes have nearly equal ionization probabilities, are inherently more noisy than silicon APDs.

The effect of avalanche multiplication can be characterized by the multiplication factor M, defined as the ratio of photocurrent with amplification to photocurrent without amplification. The detector output voltage is then still given by Eq. (14-19), with Eq. (14-18) replaced by

$$\mathcal{R} \equiv \frac{i_\lambda}{P_{\text{in}}} = \frac{Me\eta_{\text{abs}}}{h\nu} \qquad \text{(APD responsivity)} \tag{14-34}$$

M can be as high as 100 in a silicon APD, but is more typically ~ 10 in a germanium APD.

The operation of the APD depends on the proper electric field profile within the device, and this in turn requires the proper bias voltage. The relation between $E(x)$ and bias voltage (magnitudes only) is

$$\int E\, dx = V_B$$

where we have neglected the built-in potential V_0 compared with V_B. As the bias voltage increases, the area under the $E(x)$ curve increases proportionately, as illustrated in Fig. 14-15a. At some critical voltage, the depletion region "reaches through" to the highly doped p^+ region, with the electric field extending uniformly throughout the intrinsic region. Electrons generated anywhere in the intrinsic region are then efficiently swept out and injected into the high-field region for amplification. A device biased in this way is termed a *reach-through APD*.

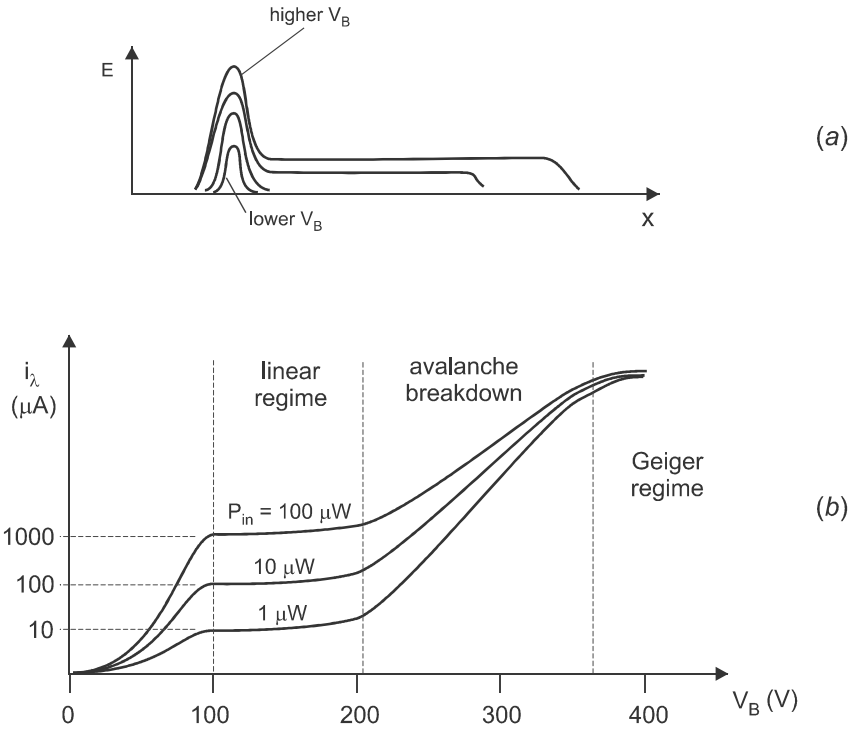

Figure 14-15 (a) The area under $E(x)$ increases with V_B until the field "reaches through" the intrinsic region. (b) Photocurrent i_λ versus reverse-bias voltage for three values of incident light power P_{in}, assuming $\mathcal{R}$ = 10 mA/mW. In the linear regime (after reach-through), $i_\lambda = \mathcal{R}P_{in}$, but in the Geiger regime i_λ becomes independent of P_{in}.

Fig. 14-15b shows a typical variation of responsivity with applied bias voltage for a reach-through APD. Below the threshold value (usually ~ 100–200 V), $\mathcal{R}$ increases with V_B as the E field starts to extend into the intrinsic region. There is a linear operating region above this, in which all photogenerated electrons are collected for amplification. In this region, the avalanche is well behaved, and the detector output is proportional to the incident light power. At still higher bias voltage, the avalanche process becomes uncontrolled, and *avalanche breakdown* ensues. In this situation, the detector output rises to a saturation value which is independent of the number of photons absorbed. This is the *Geiger mode* regime, analogous in operation to the Geiger counter used to detect nuclear radiation. The APD Geiger mode is useful when the purpose is to determine whether any photons are present, rather than to determine their number. It has applications in *photon counting,* in which the arrival time of individual photons is measured.

The responsivity of an APD generally decreases with increasing temperature, due to the increasing probability of electrons making nonionizing collisions with phonons. These collisions take away some of the kinetic energy gained by an electron that might otherwise be available for creating additional electron–hole pairs. In order to stabilize the gain of APD detectors in a changing ambient temperature, then, temperature-control circuitry is needed. Despite these complications, and the need for high-voltage bias, the improved responsivity of the APD makes it an attractive choice for applications limited by a weak light signal, such as in fiber optic communications.

Schottky Photodiode

The PIN and avalanche photodiodes discussed in the previous two sections are both p–n-junction-based devices. In contrast, the *Schottky photodiode* utilizes a metal–semiconductor junction to separate and collect the photogenerated charge carriers. Fig. 14-16 illustrates the operation of a Schottky photodiode for the metal–n–n^+ configuration, the most common type. Photons pass through a partially transparent metallic layer (often gold), and are absorbed in the n-type semiconductor. Charge carriers generated within the depletion region are efficiently swept out by the built-in electric field (see Fig. 10-17), giving rise to a photocurrent i_λ. Just as for a PIN photodiode, the diffusion tail in the time response can be minimized by adjusting the donor concentration N_D in the n-type region so that the depletion region extends all the way through to the n^+ layer. Apart from the lack of a p-type layer, the structure and operation of a Schottky photodiode is similar to that of a PIN photodiode.

Schottky photodiodes have some advantages over PIN photodiodes for certain applications. One advantage is a practical issue in manufacturing the devices. In connecting the Schottky photodiode with wires in the external circuit, only one metal–semiconductor connection needs to be made (metal–n^+), and ohmic contacts are readily formed for such a junction. Another advantage of the Schottky photodiode is an improved time response. Since it lacks the p-type layer of a PIN photodiode, there is no remnant diffusion tail arising from charge carriers generated in the p-type layer. This becomes especially important at short wavelengths, at which the large absorption coefficient would result in a significant fraction of the light being absorbed in the thin p-type layer of a PIN photodiode. Schottky photodiodes with bandwidths in the range 25–60 GHz are commercially available.

A further advantage of the Schottky photodiode is that metal junctions can be made with a wide variety of semiconductors, including those with wide band gap E_g, such as

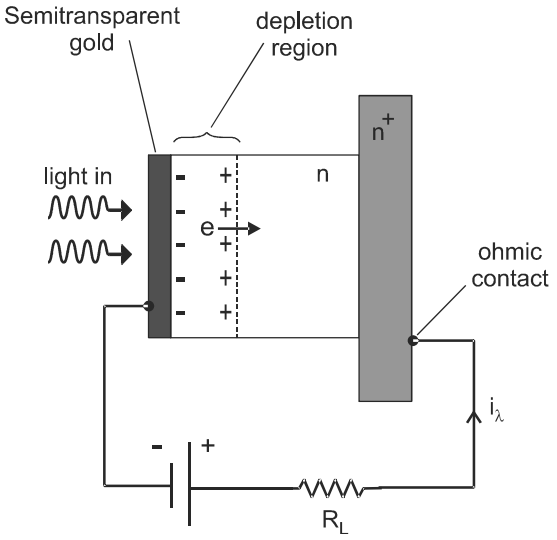

Figure 14-16 In a Schottky photodiode, light is absorbed in the depletion region of an n-type semiconductor after passing through a semitransparent metallic film.

SiC, GaN, and AlGaN. These wide-band-gap detectors have found use as "solar blind" sensors, which respond only to wavelengths in the UV region. They are inherently insensitive to sunlight in the visible and infrared regions where $h\nu < E_g$, and have applications such as flame sensors.

Schottky photodiodes do have some disadvantages, however. They tend to be less efficient than PIN photodiodes at longer wavelengths, due to reflection and absorption of light in the metal layer. To reduce light reflection, an antireflection coating is often applied, but this complicates the manufacture of such devices. Schottky photodiodes are primarily used for detecting blue or UV wavelengths, or in high-speed applications, where some loss in efficiency can be tolerated.

14-5. SIGNAL-TO-NOISE RATIO

The detectability of a small signal depends on how large it is compared to the noise. This is usually expressed by the *signal-to-noise ratio* (SNR), defined as the ratio of electrical signal power to electrical noise power. Taking the electrical signal power in the circuits of Fig. 14-2 as that due to the photocurrent i_λ, we have

$$P_{\text{sig}} = i_\lambda^2 R_L = \left(\frac{P_{\text{in}} \eta_{\text{abs}} e}{h\nu} \right)^2 R_L \qquad \text{(electrical signal power)} \tag{14-35}$$

where Eq. (14-2) has been used. Note that this expression applies only when the detector circuit is well below saturation. Also, in the photoconductive mode the "signal" current is defined as the measured current minus the dark current. The two contributions to the noise were discussed in Section 13-3. For shot noise, Eq. (13-29) gives the electrical noise power as

$$P_{\text{shot}} = (i_N)^2 R_L = 2e(i_\lambda + i_0)BR_L \qquad \text{(shot noise electrical power)} \tag{14-36}$$

where the total current $\bar{i}$ consists of both signal current i_λ and dark current i_0. For thermal noise, Eq. (13-33) gives the electrical noise power as

$$P_{\text{therm}} = \frac{V_N^2}{R_L} = 4k_B TB \qquad \text{(thermal noise electrical power)} \tag{14-37}$$

Note that the thermal noise power is independent of R_L, whereas the shot noise power increases linearly with R_L. Therefore, the dominant source of noise tends to be shot noise for large R_L, and thermal noise for small R_L.

Using the above equations, the signal-to-noise ratio can be written as

$$\text{SNR} = \frac{P_{\text{sig}}}{P_{\text{shot}} + P_{\text{therm}}} = \frac{i_\lambda^2 R_L}{2e(i_\lambda + i_0)BR_L + 4k_B TB} \tag{14-38}$$

where the signal current i_λ is related to the incident optical power P_{in} by $i_\lambda = (P_{\text{in}}/h\nu)\ \eta_{\text{abs}} e$. Since the SNR is a ratio of signal and noise powers, and the power is proportional to the square of voltage or current, we can write

$$\text{SNR} = \left(\frac{i_\lambda}{i_N} \right)^2 = \left(\frac{V_{\text{sig}}}{V_N} \right)^2 \tag{14-39}$$

where $V_{sig} = i_\lambda R_L$. It is therefore the square root of the SNR that gives the ratio of signal amplitude to rms noise amplitude.

It is useful to consider the following limiting cases, in which one source of noise dominates.

1. Large signal. When $i_\lambda \gg i_0$, and $i_\lambda R_L \gg V_T$ (recall $V_T \equiv k_B T/e$ is the voltage equivalent of temperature), we have

$$\text{SNR} \approx \frac{i_\lambda^2 R_L}{2ei_\lambda B R_L} = \frac{i_\lambda}{2eB}$$

The noise here is dominated by the shot noise from the signal current, and the SNR is independent of load resistance. Since $1/B$ corresponds to the measurement time, this result says that the SNR is roughly the number of charge carriers produced during this measurement time. Since $i_\lambda \propto P_{in}$, the SNR increases linearly with the incident optical power.

2. Small signal, large R_L. When $i_\lambda \ll i_0$ and $i_0 R_L \gg V_T$,

$$\text{SNR} \approx \frac{i_\lambda^2 R_L}{2ei_0 B R_L} = \frac{i_\lambda^2}{2ei_0 B}$$

In this regime, the SNR is limited by shot noise from the dark current i_0, and is again independent of load resistance. Note that SNR here increases with the square of the incident optical power.

3. Small signal, small R_L. When $i_\lambda \ll i_0$ and $i_0 R_L \ll V_T$,

$$\text{SNR} \approx \frac{i_\lambda^2 R_L}{4k_B T B}$$

In this regime, the SNR is limited by thermal noise from the load resistor R_L. This situation is often encountered in practice, and leads to a trade-off of SNR with detector response time. Increasing R_L improves the SNR, but degrades the response time due to RC time constant effects. Decreasing R_L improves the response time, but at a sacrifice in SNR ratio.

When the signal becomes equal to the noise (SNR = 1), it is barely discernible, and this can be considered to be the criterion for signal detectability. The optical power that gives SNR = 1 is known as the *noise equivalent power,* or NEP, and is a measure of the detector's sensitivity. In the limiting case #2 above, where shot noise from the dark current dominates, the NEP is found by setting SNR = 1 and using Eq. (14-2) with P_{in} = NEP. This gives

$$1 = \frac{(\text{NEP}\,\eta_{abs} e/h\nu)^2}{2ei_0 B}$$

or

$$\text{NEP} = \frac{h\nu}{\eta_{abs} e}\sqrt{2ei_0 B} \tag{14-40}$$

The MKS unit for NEP is watts, since it is an optical power. An alternative unit commonly used for optical power is the dBm, defined as the power in dB relative to 1 mW. Thus, for an optical power P measured in mW,

$$\text{optical power in dBm} = 10 \log_{10}\left(\frac{P}{1\ \text{mW}}\right) \tag{14-41}$$

For example, an optical power of –20 dBm is 0.01 mW, whereas an optical power of +20 dBm is 100 mW.

It is useful to separate the NEP for a detector into those factors that are fundamental and those that can be adjusted by the device geometry or detector circuit. The dark current i_0, for example, is not fundamental, since it is proportional to the area of the p–n junction. The fundamental parameter is the dark current density J_0, which depends on the material used and the temperature, but not on the device geometry. Writing $i_0 = J_0 A$, we obtain

$$\text{NEP} = \frac{h\nu}{\eta_{\text{abs}}}\sqrt{\frac{2J_0 AB}{e}} \tag{14-42}$$

This shows that the minimum power that can be detected is proportional to $\sqrt{AB}$. The detector can be made more sensitive by decreasing the junction area A, or by decreasing the detection circuit bandwidth. To obtain a parameter that is independent of B and A, the NEP can be divided by $\sqrt{AB}$. It is conventional to define the reciprical of this as a figure of merit, since it is then a larger number for a better (more sensitive) detector. Designating this figure of merit as D^* (pronounced "dee star"), we have

$$D^* \equiv \frac{\sqrt{AB}}{\text{NEP}} \tag{14-43}$$

which becomes

$$\begin{aligned} D^* &= \frac{\eta_{\text{abs}}}{h\nu}\sqrt{\frac{e}{2J_0}} \\ &= \frac{\mathcal{R}}{\sqrt{2eJ_0}} \end{aligned} \tag{14-44}$$

where Eq. (14-18) was used in the last step.

The D^* parameter provides a good way of comparing the ultimate sensitivity limits for different types of detectors. According to Eq. (14-44), D^* varies inversely as the square root of dark current density, so that narrow-band-gap materials (with high J_0) have a smaller D^* than wider-band-gap materials. Since a narrower band gap is required to detect light of longer-wavelength, D^* is inherently smaller for longer wavelength detectors, all other things being equal. D^* also depends on wavelength through the responsivity $\mathcal{R}(\lambda)$. Fig. 14-17 shows the wavelength dependence of D^* for some common detector materials. Note that since J_0 decreases with decreasing temperature, the D^* for a given photodetector can generally be improved by lowering the temperature of the semiconductor element.

14-6. DETECTOR CIRCUITS

So far, we have considered only the simple detector circuits shown in Fig. 14-2, in which the output is taken as the voltage across the series load resistor R_L. Here, we consider two types of detector circuits that provide not only the proper bias for the diode, but also a degree of amplification.

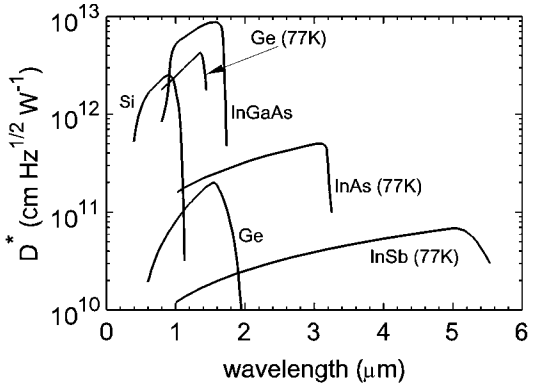

Figure 14-17 The specific detectivity parameter D^* for some respresentative photodetector materials. The maximum possible D^* decreases at longer wavelength because the dark current is higher for narrower-band-gap semiconductors.

High-Impedance Amplifier

Figure 14-18 shows one scheme for amplifying the signal in a photodiode circuit. This is basically the same as the circuits of Fig. 14-2, except that an additional amplification stage has been added with an FET (field effect transistor). The voltage generated across the load resistor is applied between the gate (G) and source (S) of the FET, and this results in an amplified output voltage between the source and drain (D). For the best possible SNR, the photodiode can be operated in the photovoltaic mode, where $V_B = 0$ and there is no dark current. The signal output will be proportional to R_L (below saturation), so higher R_L is best for detecting very small signals. This circuit is a good choice when the best possible SNR is desired.

If time response is important, however, this is not the best circuit to use. The large load resistance, in combination with the diode capacitance C_{diode}, gives a 3 dB electrical bandwidth:

$$B = \frac{1}{2\pi R_L C_{\text{diode}}} \qquad \text{(high-impedance amplifier bandwidth)} \qquad (14\text{-}45)$$

This circuit, then, suffers from the same sensitivity/time response trade-off that we discussed earlier.

Transimpedance Amplifier

When response time is important, a better choice for detector circuit is the one shown in Fig. 14-19. This circuit uses an operational amplifier (op-amp) to convert the photocurrent i_λ directly into an output voltage, hence the term *transimpedance amplifier.* The op-amp has the property that the two input terminals are held at nearly the same potential (virtual ground), while at the same time very little current is allowed to flow into or out of either terminal. For the purpose of biasing the photodiode, then, the op-amp input acts like a short circuit ($R_L = 0$), which keeps the diode below saturation for any level of light input. Any photocurrent must flow not through the input terminals of the op-amp, but

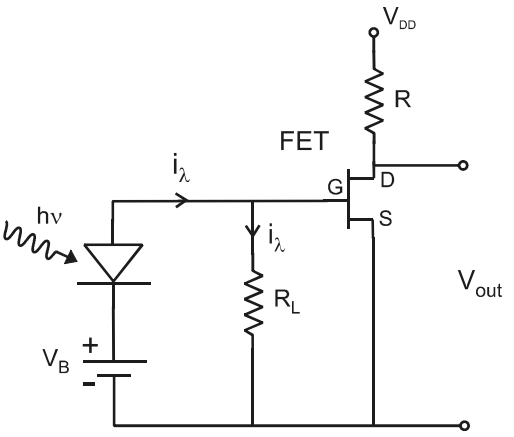

Figure 14-18 High-impedance FET amplifier circuit for the photodiode.

rather through the feedback resistor R_F, which is connected between the input and output of the op-amp. The output voltage is then equal to the voltage across this feedback resistor, $V_{\text{out}} = i_\lambda R_F$, since both input terminals of the op-amp are at ground potential. The circuit acts as if $R_L = R_F$ in terms of the output voltage, but it acts as if $R_L = 0$ in terms of biasing the photodiode. This gives the best possible linearity and dynamic range for the photodiode, and the output voltage is in fact limited only by the maximum output voltage of the op-amp (typically ~ 10 V).

A further advantage of the transimpedance amplifier is seen in the time response. Since the diode voltage is held essentially constant, the capacitance of the diode's p–n junction no longer limits the time response. Instead, it is the feedback capacitance C_F characteristic of the op-amp that matters, and the bandwidth becomes

$$B = \frac{1}{2\pi R_F C_F} \qquad \text{(transimpedance amplifier bandwidth)} \tag{14-46}$$

Since C_F can be much smaller than C_{diode}, the transimpedance amplifier can have a much higher bandwidth for the same sensitivity ($R_F = R_L$). This type of detector circuit is the preferred one in many situations, because of the advantages of high speed and large dynamic range. The only drawback is in obtaining the best possible SNR for weak signals, in which case a photodiode in the photovoltaic mode followed by a high-impedance amplifier is the best choice.

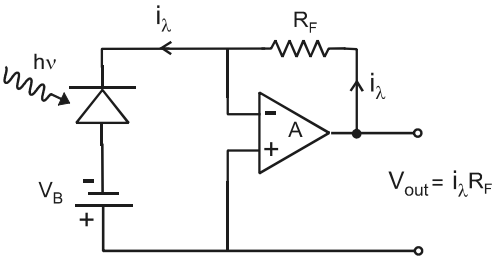

Figure 14-19 Transimpedance amplifier circuit for the photodiode.

PROBLEMS

14.1 A photodiode with responsivity 0.3 A/W and dark current 2 nA is biased in the photoconductive mode, with a 9 V battery and 500 kΩ resistor. Make a sketch like that of Fig. 14-3, showing the load line and the diode curves for incident powers from zero to 100 μW in steps of 20 μW. Circle the operating point for an incident power of 40 μW, and determine the approximate diode voltage from the graph.

14.2 For Problem 14.1, make a sketch of the output voltage (across the resistor) versus the incident optical power, for the range 0 to 100 μW. At what optical power does the detector response saturate?

14.3 The photodiode of Problem 14.1 is removed from the circuit and operated in the photovoltaic mode. (a) Determine the shunt resistance assuming $\beta = 2$. (b) Under open circuit conditions (no load resistor), what incident optical power will result in saturation of the output voltage? (c) A load resistor is now added to increase the dynamic range. What value of load resistance is needed so that optical powers up to 20 μW can be detected without saturation?

14.4 Show that Eq. (14-13) can be obtained from the equivalent circuit model shown in Fig. 14-6.

14.5 A silicon solar cell has area 50 cm^2, reverse-saturation current 0.75 μA, $\beta = 2$. The electrical power generated in the 0.4 Ω load resistor is 894 mW. Determine (a) The circuit current i, (b) the photocurrent i_λ, (c) the optical power incident on the cell, assuming that 80% is absorbed, and (d) the optical-to-electrical conversion efficiency of the cell. Assume the temperature remains near 300 K. Assume $\lambda = 500$ nm for the incident light.

14.6 Using Eq. (14-26) for an exponential voltage rise, show that the rise time (10% to 90% points) is given by $t_r \simeq 2.2RC$.

14.7 A silicon p–n junction photodiode has junction area 1 cm^2, and doping levels 10^{14} and 10^{16} cm^{-3} on the n and p sides, respectively. It is reverse biased with 15 V and a 10 kΩ load resistor is used. (a) Determine the 3 dB electrical bandwidth due to the RC time constant. (b) Determine the bandwidth due to the hole transit time. (c) What is the limiting bandwidth in this case?

14.8 Assume the total response time of a silicon photodiode can be taken as the sum of the transit time (limited by saturation velocity 10^5 m/s) and the RC rise time t_r. Derive an expression for the optimum intrinsic region thickness d. If the load resistance is 50 Ω and the detector area is 0.01 mm^2, calculate d and the resulting detector bandwidth.

14.9 A high-speed germanium PIN photodiode has a depletion width of 10 μm and a reverse-bias voltage of 10 V. The hole mobility in Ge is $\simeq 0.2$ $m^2/(Vs)$, the saturation velocity is $\simeq 7 \times 10^4$ m/s, and the refractive index at 1300 nm is $\simeq 4.3$. (a) Determine the transit time limit to the response time, and calculate the corresponding 3 dB electrical bandwidth. (b) If light of wavelength 1300 nm is detected, determine the fraction of incident light that is absorbed in the depletion region (include the reflection loss from the air–Ge interface). (c) Repeat part b if the detected wavelength is 1600 nm. See Fig. 13-16 for Ge absorption coefficient.

14.10 A silicon APD has a responsivity of 20 A/W at the detection wavelength of 850 nm, and the absorption efficiency is 0.7. Determine the avalanche gain.

14.11 A silicon photodiode is configured as shown in Fig. 14-18 with a 90 V bias voltage. The light to be detected has intensity 20 $\mu W/cm^2$ and wavelength 920 nm. Relevant material properties for the detector are: absorption efficiency = 0.18, dark current density at room temperature = 15 nA/cm^2, charge carrier mobility = 0.048 m^2/Vs, and carrier saturation velocity = 10^5 m/s. At the applied bias voltage, it is known that the width of the depletion region is 0.2 mm. (a) If the photocurrent is 150 nA, what is the area of the detector? (b) If the load resistor is 100 kΩ, determine the RC time constant of the circuit, and the corresponding 3 dB bandwidth. Video requires a bandwidth of about 2 MHz. Will the circuit be suitable for video applications? (c) Determine the transit-time response for the circuit, and compare it with the RC time constant. Which is the primary limit to the bandwidth in this circuit? (d) Repeat part b assuming a load resistance of 10 kΩ. Is the circuit now suitable for video applications?

14.12 In Problem 14.11, check to see that there is sufficient signal-to-noise (S/N) ratio. Using the results for the 10 kΩ load resistor, determine the power S/N ratio, and also find the ratio of the rms deviation in signal voltage to the average signal voltage (express as a percentage).

14.13 The photodetector of Problem 14.11 is now used for low-level dc light-level measurements. Assume that in this application the effective bandwidth is 1 Hz. (a) Determine how large the load resistor must be in order for the noise to be dominated by dark current shot noise rather than by thermal noise. Take as the criterion that the shot noise power is five times the thermal noise power. (b) In the limiting case described in part a, determine the noise equivalent power (NEP) for the detector (in units of watts). (c) For the conditions described in part a, determine the minimum light intensity that can be detected with this detector, taking as the criterion that the signal voltage must be 10 times the rms noise voltage.

14.14 Consider the transimpedance amplifier optical receiver shown in Fig. 14-19. The feedback resistance is 10 kΩ and the feedback capacitance is 0.2 pF. The diode's capacitance is 5 pF, and its responsivity is 0.5 A/W. The incident optical power is 0.5 mW. (a) Compute the signal current. (b) Compute the receiver's output voltage. (c) Compute the receiver's 3 dB electrical bandwidth. (d) Compute the rms thermal-noise current generated in the feedback resistor, assuming a temperature of 300K. (e) Assuming no dark current, and an ideal (noiseless) amplifier, compute the output SNR, expressed in dB. The actual SNR will be somewhat lower due to noise introduced by the amplifier.

14.15 Use the data in Fig. 14-17 to determine the following: (a) The minimum optical power at 900 nm that can be detected (SNR = 1) by a Si photodiode of area 0.02 cm^2 in a 1 Hz bandwidth, (b) the dark-current density of an InAs detector at 77 K, assuming that the absorption efficiency is near unity for λ = 2.8 μm, and (c) the dark current and minimum detectable power for an InGaAs detector of area 0.02 cm^2, operating at 1550 nm in a 1 Hz bandwidth. Assume an absorption efficiency near unity at 1550 nm.

Samsung Exhibit 1044, Page 34 of 34